MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo #377 August 1976

REPRESENTATION AND RECOGNITION OF THE SPATIAL ORGANIZATION OF THREE DIMENSIONAL SHAPES

by

D. Marr and H. K. Nishihara

ABSTRACT. A method is given for representing 3-D shapes. It is based on a hierarchy of stick figures (called *3-D models*), where each stick corresponds to an axis in the shape's generalized cone representation. Although the representation of a complete shape may contain many stick figures at different levels of detail, only one stick figure is examined at a time while the representation is being used to interpret an image. By thus balancing scope of description against detail, the complexity of the computations needed to support the representation is minimized. The method requires (a) a database of stored stick figures; (b) a simple device called the *image-space processor* for moving between object-centered and viewer-centered coordinate frames; and (c) a process for "relaxing" a stored model onto the image during recognition. The relation of the theory to "mental rotation" phenomena is discussed, and some critical experimental predictions are made.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643.

Summary

1. A method is given for representing 3-D shapes. It is based on a hierarchy of stick figures (called *3-D models*), where each stick corresponds to an axis in the shape's generalized cone representation.

2. By using stick figures to represent a shape and its parts at several levels of detail, a representation is obtained that is intrinsically simple, yet which maintains its fidelity to an arbitrary level of precision.

3. While the representation is being used to interpret an image, only one stick figure is examined at a time. By thus balancing scope of description against detail, the complexity of the computations needed to support the representation is minimized.

4. The structures and processes associated with the method are described. The most important are (a) a database of stored stick-figures, which are indexed in several ways; (b) an *image-space processor*, which is a simple mechanism for moving between object-centered and viewer-centered coordinate frames; and (c) a process for "relaxing" a stored model onto the image during the recognition and representation of spatial orientation.

5. Some facets of the theory's relaxation process resemble the computation of a 3-D rotation, but a computer graphics metaphor is misleading. In fact the manipulations take place on abstract vectors (the sticks) that are not even present in the original image, and it is roughly correct to say that only two such vectors are explicitly represented at a time.

6. If the method is taken as a psychological theory, it makes a critical prediction which, if false, would disprove it. Views of an object in which an important axis of its generalized cone representation is severely foreshortened are peculiarly difficult to interpret. Such views are not uncommon, and it is predicted that this class of views corresponds to those that Warrington & Taylor (1973) labelled "unconventional". Their patients should therefore fail on these views.

7. The theory provides an explanation of most of the experimental results concerning mental rotation that have recently been discovered by R. N. Shepard and his colleagues. The linear dependence between time to interpret and 3-D angular discrepancy is however not a deep consequence of the theory, merely the signature of implementing it in a particularly simple way.

I: Introduction

At some point during the analysis of a two dimensional image of an object, the three-dimensional structure of the viewed object and its spatial relation to the viewer must be established and represented. The question is how? The form of the answer we require is not a detailed specification of some complex neurophysiological mechanism, although eventually one will wish to derive such a thing. First, we need a more abstract understanding of the computational problems involved, that shows when and how to use the various kinds of information that are available from an image. The understanding that we seek may be expressed as a *method* (see Marr 1976a); it amounts to a competence theory for this aspect of 3-D vision.

This article presents such a method, and it has four key ingredients:

(a) The deep structure of the three-dimensional representation of an object's shape consists of a coarse stick figure, whose sticks correspond to axes of the major components of the shape (such as arms, torso, head); and of individually addressable stick figures for each of the component shapes. In this way, arbitrary detail can be represented in a system each of whose component stick figures is rather simple, yet which maintains faithfully the important shape characteristics at each level of description.

(b) Each stick figure is defined by a propositional database called a *3-D model*. The geometrical structure of a 3-D model is specified by storing the relative orientations of pairs of connecting sticks. Thus the specification is made in a local coordinate system based on the principal component of the shape at that level of description, not in absolute coordinates based on a circumscribing frame of reference.

(c) When a 3-D model is being used to interpret an image, a computation must be made that relates the geometrical relationships among the sticks of the 3-D model to the 2-D relationships among the projections of those sticks in the image. The computation depends upon the orientation and location of the 3-D model relative to the viewer. This is accomplished by a computationally simple mechanism called the *image-space processor*, which may be thought of as a device for transforming a vector between object-centered and viewer-centered coordinate systems.

(d) During recognition, a sophisticated interaction takes place between the image, the 3-D model, and the image-space processor. This interaction gradually relaxes the stored 3-D model so that its axes project onto the axes computed from the image. Some facets of this process resemble the computation of a 3-D rotation, but a simple computer graphics metaphor is misleading. In fact, the rotations take place on abstract vectors (the axes) that are not even present in the original image, and it is roughly correct to say that only two such vectors are explicitly represented at a time.

Thus the essence of the theory is a method for representing the spatial disposition of the parts of an object and their relation to the viewer. We believe that it may shed some light on the phenomena of mental rotation uncovered by R. N. Shepard and his collaborators, and on certain neurological findings reported by Warrington & Taylor (1973).

Background: modular decomposition of the recognition process

Our overall picture of the recognition problem is illustrated in figure 1, which embodies two points that we take as assumptions. Firstly, to a first approximation the

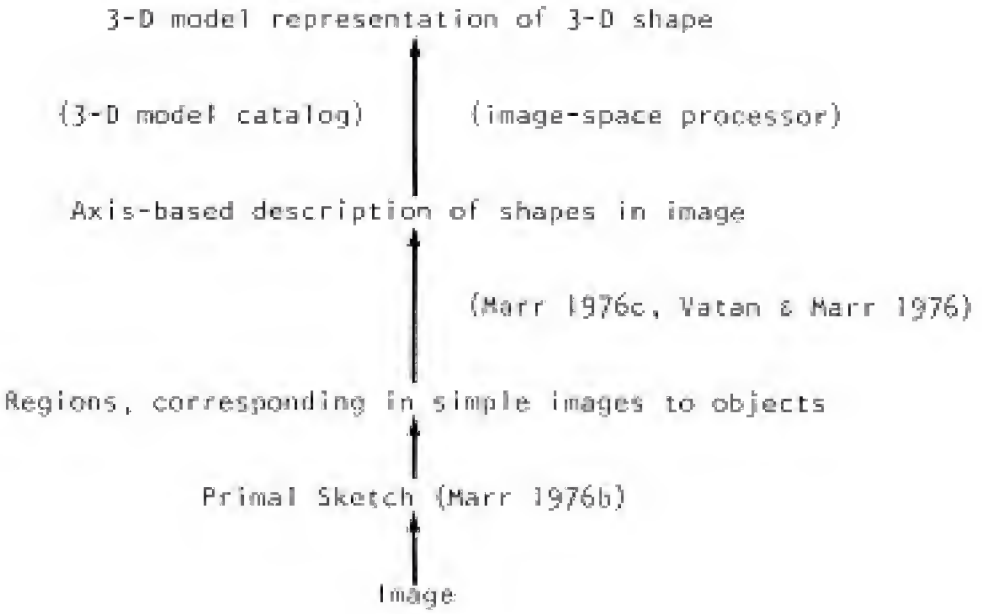


Figure 1. This diagram summarizes our overall view of the visual recognition problem, and it embodies several points that this article takes as assumptions. The first is that the recognition process decomposes to a set of modules that are to a first approximation independent. The simplified subdivision shown here consists of four main stages, each of which may contain several modules. (1) The translation of the image into a primitive description called the *primal sketch* (Marr 1976b); (2) The division of the primal sketch into regions or forms, through the action of various grouping processes ranging in scope from the very local to global predicates like a rough type of connectedness; (3) The assignment of an axis-based description to each form (see figure 4); and (4) The construction of a 3-D model for the viewed shape, based initially on the axes delivered by (3). The relation between the 3-D model representation of a shape and the image of that shape is found and maintained with the help of the image-space processor. Finally, the representation of the geometry of a shape is separate from the representation of the shape's use or purpose (Warrington & Taylor 1973).

process of visual recognition decomposes to a set of modular steps. The evidence for this is extensive but indirect. It includes evidence from electrophysiological recordings from Adrian (1941) to Hubel & Wiesel (1962, 1965), Barlow, Blakemore & Pettigrew (1967), and Zeki (1973); histological and neuroanatomical evidence from Brodmann (1909) and Cajal (1911) to modern studies such as those of Zeki (1971), Allman, Kaas, Lane & Miezin (1972), Allman, Kaas & Lane (1973), Allman & Kaas (1974a, b & c) and the mass of clinical studies describing patients who have lost particular and highly circumscribed functional parts of their perceptual or motor faculties (Critchley 1953, Luria 1970, Vinken & Bruyn 1969). Evidence against the assumption of modularity in its strictest form comes from illusions in which quite late processing or high-level knowledge about an image appears to influence earlier processing; for example, shape recognition normally follows figure-ground separation, but can sometimes influence it (e.g. Street 1931). According to the assumption of modularity, these effects should be regarded as second-order interactions between modules that are to a first approximation independent (Marr 1976b).

Secondly, we assume that there exists a module (or group of modules) that is concerned with describing the 3-D shape of an item, and that this module is separate from the representation of an item's functional semantics. The evidence for this is a penetrating analysis by Warrington & Taylor (1973), who concluded that these two functions reside in distinct cortical areas. Patients with left parietal lesions showed disorders related to the use and purpose of an object, but their ability to recognize and represent its 3-D shape appeared to be intact. The opposite was true of patients with right parietal lesions.

This article describes a theory of the representation and recognition of 3-D shape. Some parts of it, including the 3-D representation scheme and the image-space processor, are precisely defined. Other parts, for example those concerned with database access during recognition, are not yet rigorous. The reader will recognize that the looser parts of the theory are those that are closely intertwined with other modules that we have not yet studied, and cannot be made precise until the exact nature of those modules, and what they can deliver from an image, has been defined. We recognize the shortcomings in this account that arise for this reason, but believe that in order to rectify them one has to have a clear grasp of a larger portion of the overall recognition process than the particular 3-D module described here. The theory as described here, together with other work (Marr 1976b, Marr 1976c, Vatan & Marr 1976, Marr & Poggio 1976a, Ullman 1976) summarised briefly by Marr & Poggio (1976b), represents an attempt at decomposing the vision problem into modules. Study of the interactions between modules must follow this.

General nature of the 3-D representation

Methods for deriving and manipulating the representation of a 3-D shape depend heavily on the nature of the representation used. Our first task is therefore to discover which representation is most appropriate. There are four ideas current in the literature; the "multiple view" representation described by Minsky (1975), Baumgart's (1975) representation by polyhedral approximation, the "generalized cylinder" representation proposed by Binford (1971), and Blum's (1973) "symmetric axis" representation, which is similar to the generalized cylinder representation for 2-D shapes, but differs from it in three

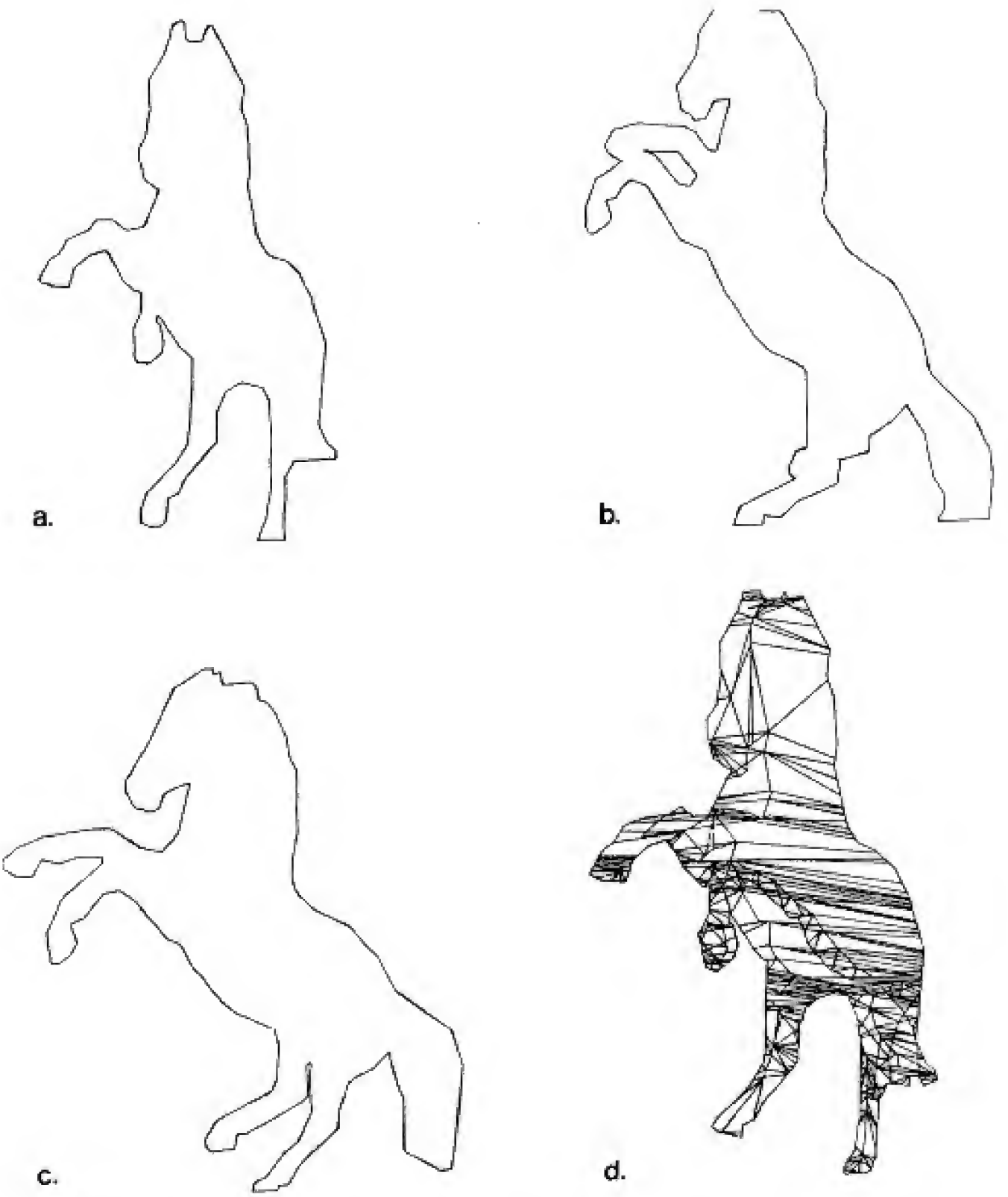


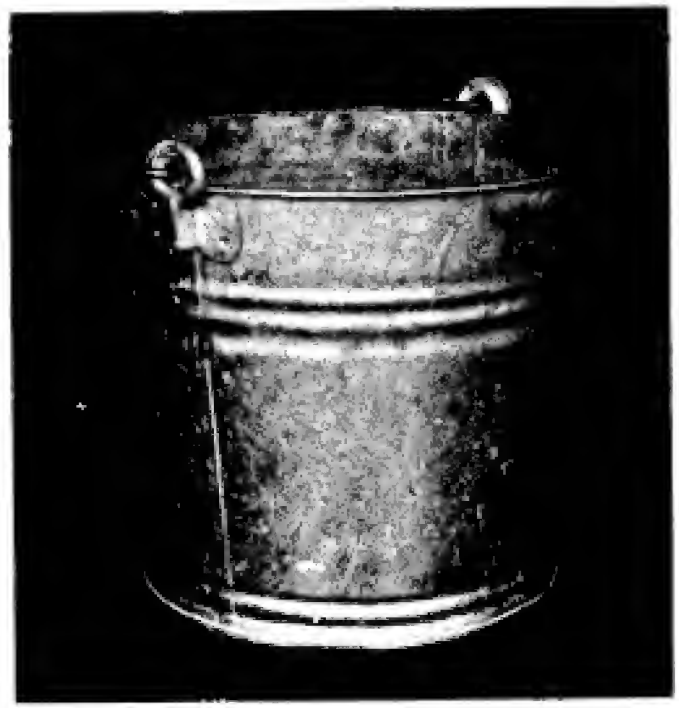
Figure 3. This figure is taken from figure 1 of Baumgart (1975), and illustrates his representation of 3-D shape by polyhedral approximation. From three views of a plastic horse, the silhouettes (a), (b) and (c) were obtained. A 3-D structure was computed from these silhouettes by a cone intersection technique, and the polyhedral representation of the resulting shape is illustrated in (d). Various disadvantages of this representation, of which the most severe is its lack of uniqueness, combine to render it an unlikely candidate for the psychological representation of 3-D shape.

dimensions.

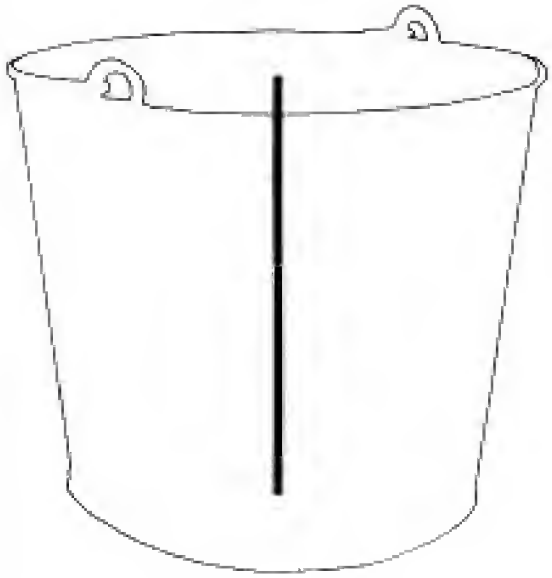
The multiple view representation is based on the insight that if one chooses one's primitives correctly (e.g. the "side" of a cube), the number of qualitatively different views of an object may be quite small. Minsky (1975) proposed that the representation of a 3-D shape might therefore consist of a catalogue of the different appearances of that shape, and that this catalogue would not need to be too large. The multiple view representation is at present underdefined -- for example, are all "views" of a man the same in which the same limbs are visible but arranged in different positions? -- and so it is difficult to argue cogently against it. Nevertheless something of a case against it can be made from Warrington & Taylor's (1973) findings. The side view of a water pail is very different from the top view, and both are reasonably simple (see figure 2). Since both views are probably equally common, one would expect the multiple view representation to contain and (presumably) to have indexed both of them. If the lesions of Warrington & Taylor's patients had randomly damaged a multiple view representation, one would expect some patients to have lost one view, and others, another. But the finding is that all patients are impaired on the same view (the one from above), views that Warrington & Taylor called "unconventional". Although the multiple view representation is not absolutely incompatible with these findings, strong extra assumptions are needed to incorporate them.

Baumgart (1975) has proposed using a system of polyhedral approximations to 3-D shapes (see figure 3). The motivation for this is that computer graphics systems make it easy to manipulate representations constructed of straight-edge segments, and the comparison between the expected view and the actual view of a polyhedral structure is therefore feasible. He makes no claims that this representation has any psychological importance, however, and the features that make it attractive for machine vision tend to make it an unattractive candidate for psychology. Although Baumgart has addressed with some success the problem of constructing a 3-D model from several views of an object, he has not shown how to recognize a known model from just one monocular view. More seriously, there is no real sense of uniqueness in his representation. A horse shape can be approximated in many ways by polyhedra, and there is no guarantee that the representations obtained on two different occasions from different sets of views will be homologous. A representation that lacks a strong uniqueness condition will be almost useless for recognition. There are also other difficulties with polyhedral approximation. They include the lack of any natural representation of articulation of parts of an object (e.g. arms and legs); the difficulty of answering overall questions about an object, like where it is pointing, given only a set of polyhedra each of which describes some small part; and the complex way in which joins between polyhedra have to be specified. As a candidate for psychology, this representation at present seems to have no particular advantages and several disadvantages. We shall therefore not consider it further.

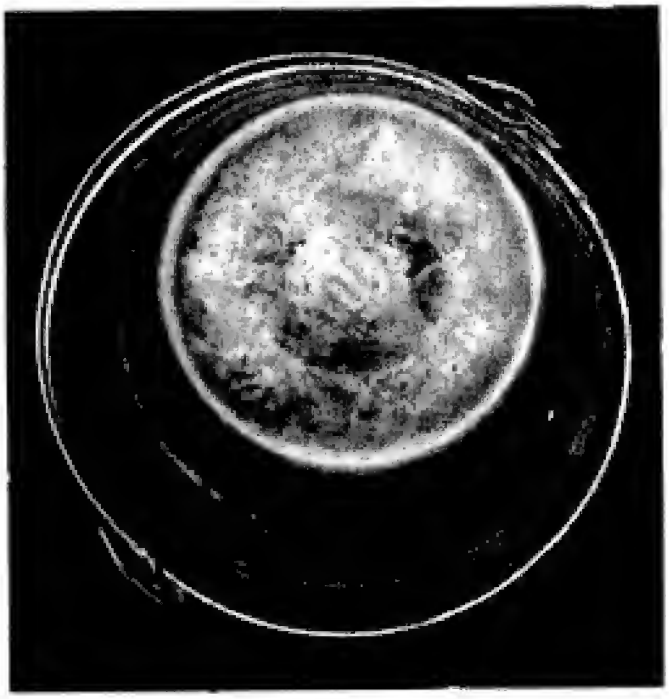
A *generalized cylinder* is the surface swept out by moving a cross-section along an axis. The axis need not be straight, and the cross-section may vary. The generalized cylinder representation of an object is obtained by splitting it up into components each of which is described in this way. A *generalized cone* is a generalized cylinder in which the shape of the cross-section remains constant but for smooth variations



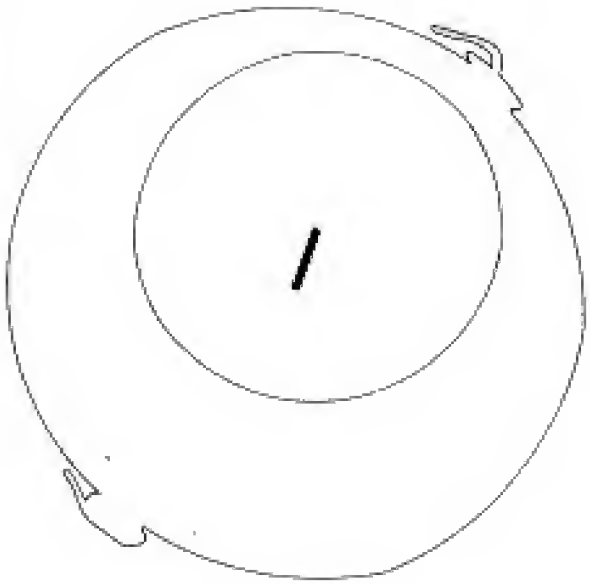
a.



c.



b.



d.

Figure 2. (a) and (b) show two views of a water-pail. Warrington & Taylor's (1973) patients are impaired on (b), but not on (a). This is difficult to reconcile with Minsky's (1975) multiple view representation, since both views are about as common. It is consistent with the 3-D model representation, for reasons that are clear from (c) and (d). The outlines of the original figures are shown as thin lines, and the axis is shown as a thick one. This axis is directly recoverable from image (a), but not from (b) where it is severely foreshortened. Since the 3-D model representation relies on an explicit representation of this axis, the successful recognition of views like (b) requires considerable extra computation.

in size.

Agin (1973) and Nevatia (1974) used a laser range-finding technique to obtain the generalized cylinder representation of objects such as a Barbie doll, a snake and a horse. Hollerbach (1975) showed how contour information may be used to derive the generalized cylinder representation of a wide range of pottery, and he found that the descriptive terminology for such artifacts in the archaeological literature corresponds naturally to terms that appear in the generalized cylinder representation. Marr (1976c) has proved that certain assumptions, which are implicit in the derivation of shape from contour, are equivalent to assuming that the viewed shapes are composed of generalized cones; and Vatan & Marr (1976) have constructed algorithms for segmenting the monocular image of a shape into its generalized cone components (see figure 4).

Blum (1973) has developed a geometry of shape based on the notion of growth outward from a point. In two dimensions, his representation may be obtained by imagining a fire lit at all points around an outline. The fire from opposite "sides" of a figure will meet in the middle, along what Blum calls the figure's "symmetric axis". The representation consists of inverting this process, specifying the symmetric axis and the degree of growth outward from each point on it.

For two-dimensional shapes, this representation resembles the generalized cylinder representation, although it is not identical. For three dimensions however, the "symmetric axis" may be two-dimensional, so this representation differs from generalized cylinders in a substantial way. Of the two representations, generalized cones seem to be preferable because for three-dimensional surfaces they are simpler, and because of their intimate connection with assumptions that are implicit in the interpretation of occluding contours in an image (Marr 1976c).

The generalized cone representation introduces two main problems; obtaining the axes and the cross-sections of the different parts of an object (arms, legs, torso), and representing the spatial disposition of the components thus obtained. These tasks are nearly independent, and this article is concerned only with the second of them, how to represent the arrangement in space of the different cones into which the viewed shape is decomposed. To solve this problem, it is enough to represent the spatial dispositions of the *axes* that occur in an object's generalized cone representation, which is equivalent to the problem of describing stick figures -- models made out of pipe-cleaners, one for each axis (see figure 5). Such models exhibit *only* the lengths and disposition of axes in the generalized cylinder representation, yet we can easily discern the giraffe, ostrich and goat in the figure. That their recognition is so easy makes it reasonable to suppose that we ourselves decompose the 3-D representation problem into similar components.

II: The Structures of the theory

The theory consists of a method for determining and representing the three-dimensional dispositions of a stick figure's axes for the purpose of recognition, given only a two-dimensional projection of those axes. It rests on the interplay between the image and two other structures: a database of stored representations of shapes (the 3-D models), and a mechanism for performing coordinate transforms (the image-space processor). The

Figure 4. Analysis of a contour from Vatan and Marr (1976). The outline (a) was obtained by applying local grouping operations to a primal sketch (Marr 1976b). It is then smoothed, and divided into convex and concave components (b). The outline is searched for deeply concave points or components, which correspond to strong segmentation points. One such point is marked with an open circle in (c). There are usually several possible matching points for each strong segmentation point, and the candidates for the marked point are shown here by filled circles (c). The correct mates for each segmentation point can usually be found by eliminating relatively poor candidates. The result of doing this here is the segmentation shown in (d). Once these segments have been defined, their corresponding axes (thick lines) are easy to obtain (e). They do not usually connect, but may be related to one another by intermediate lines which are called *embedding relations* (thin lines in f). According to the present theory, the resulting stick figure (f) is the deep structure on which interpretation of this image is based.

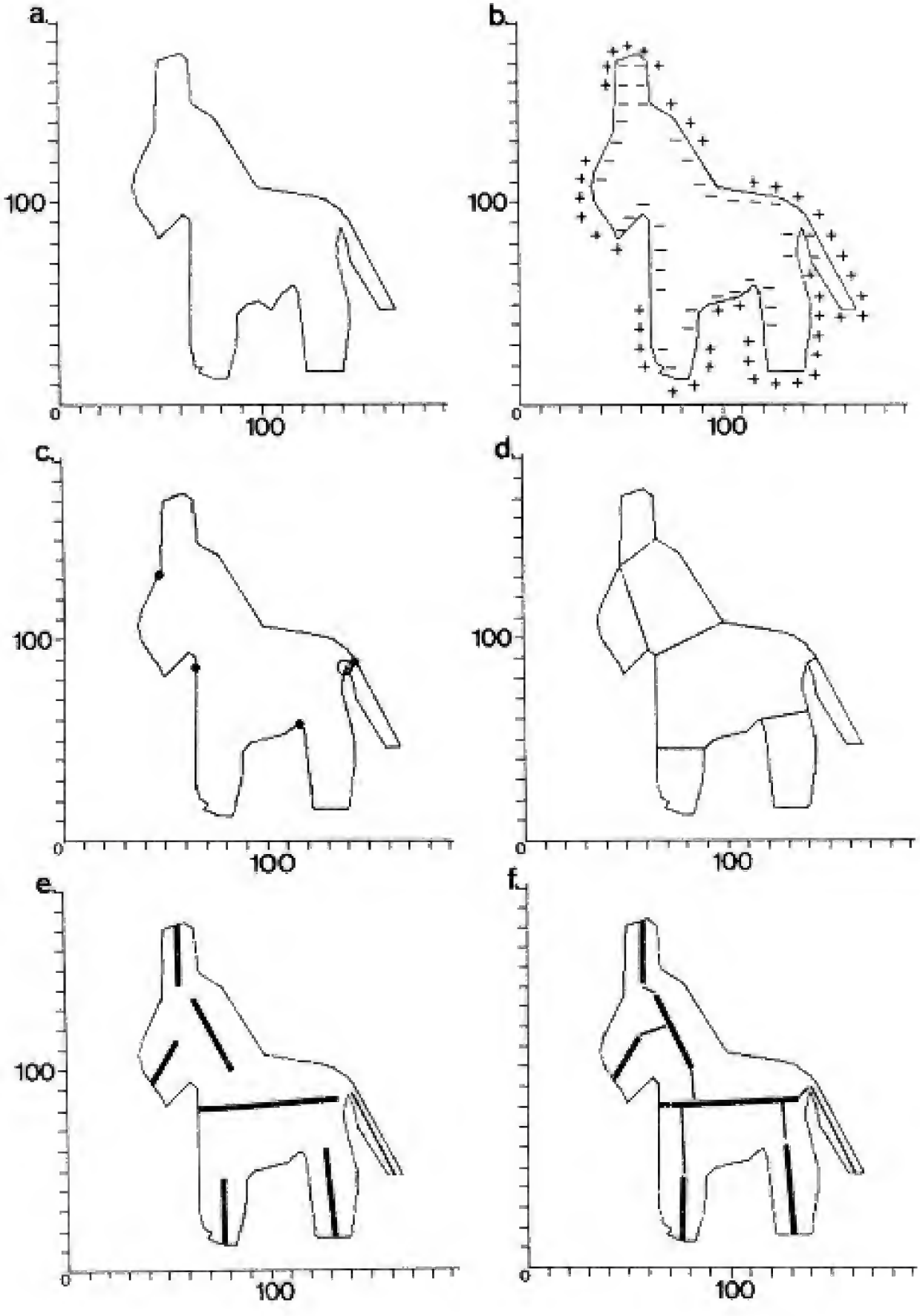

a.
b.
c.
d.
e.
f.
100
0
100

Figure 5. The theory asserts that the 3-D representation of a shape is decomposed into two parts, the description of the cross-sections that occur in the shape's generalised cone representation, and the disposition of the axes of these cones in space. Our theory deals with the second problem, which is essentially the problem of describing stick figures. The shapes in these pictures were made out of pipe-cleaners. The reader will have no trouble in recognising the giraffe, goat, rabbit and ostrich. That their recognition is so easy makes it reasonable to suppose that at some stage, we ourselves decompose the 3-D representation problem into similar components.

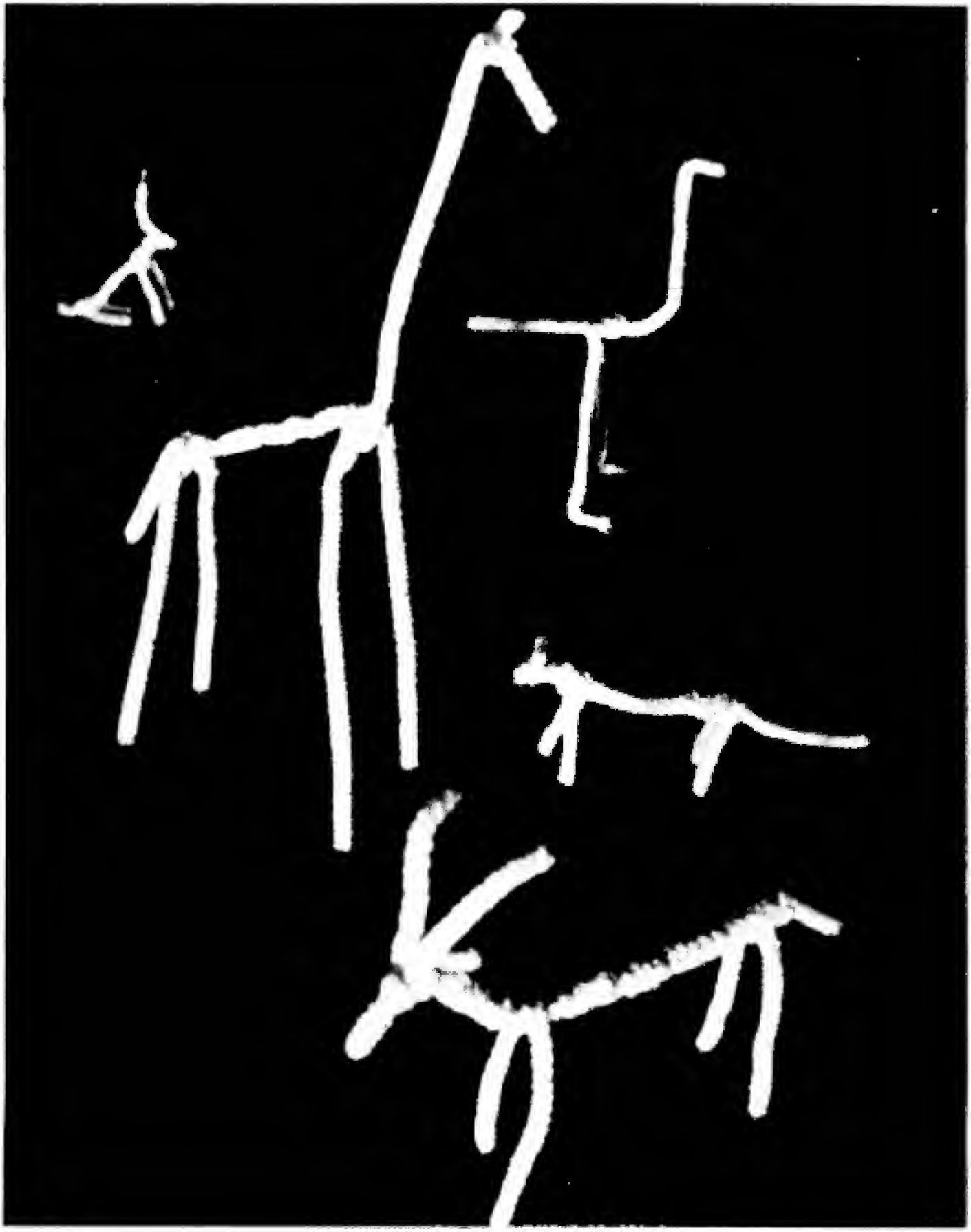
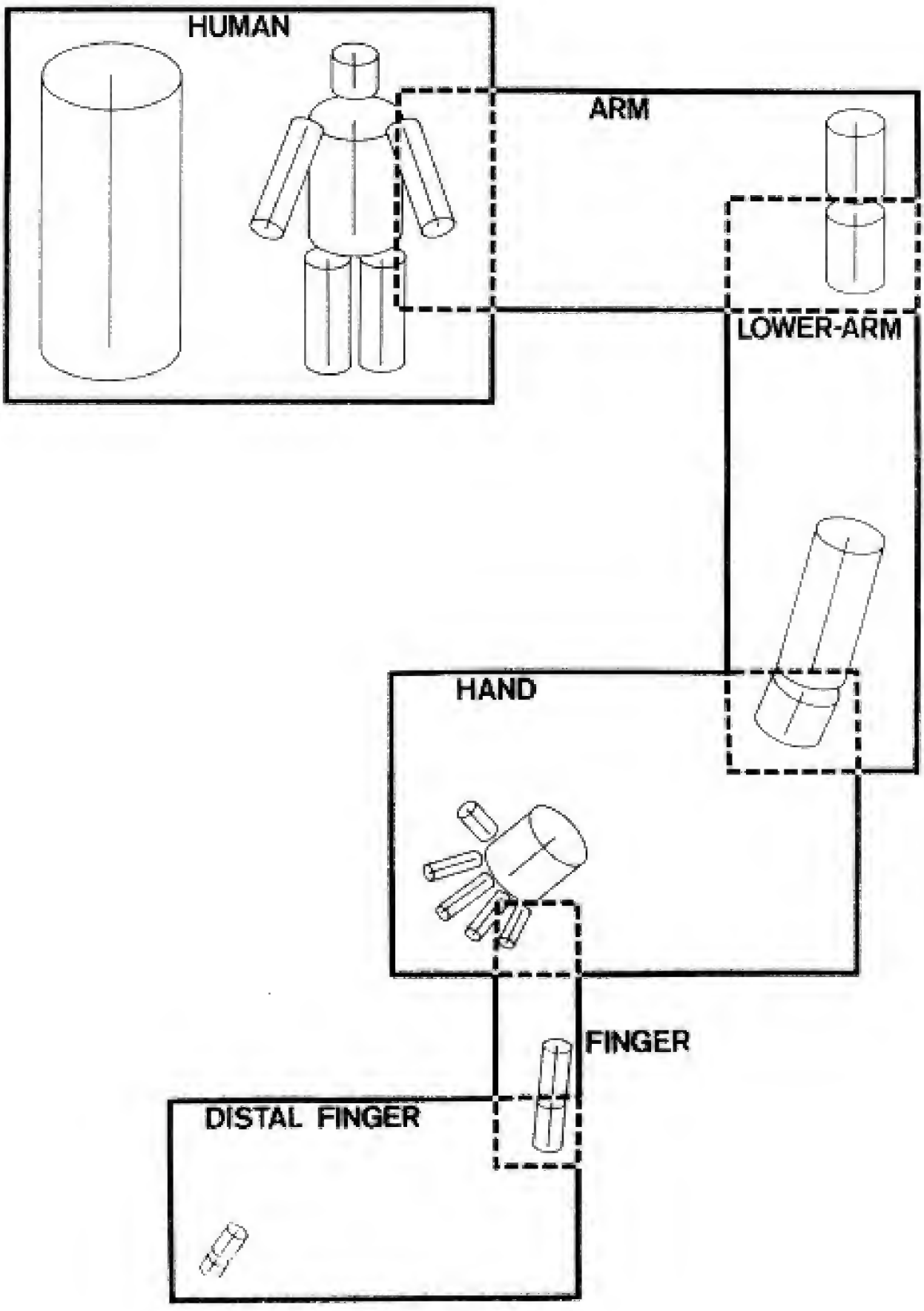


Figure 6. Examples of 3-D models, and their arrangement into the 3-D model representation of a human shape. A 3-D model consists of a model axis and component axes (left and right figures respectively in box labeled HUMAN) the latter consisting of a principal axis (the torso) and several auxiliary axes (the head and limbs) whose positions are described relative to the principal axis. The complete human 3-D model is enclosed in the rectangle labeled HUMAN. The 3-D model representation is obtained by concatenating 3-D models for different parts at different levels of detail. This is achieved by allowing a component axis of one 3-D model to be the model axis of another. Here, for example, the arm auxiliary axis in the human 3-D model acts as the model axis for the arm 3-D model, which itself has two component axes, the upper and lower arms. The figure shows how this scheme extends downwards as far as the fingers.


HUMAN
ARM
LOWER-ARM
HAND
FINGER
DISTAL FINGER

basic strategy of our approach rests on the principles of least commitment and graceful degradation (see below and Marr 1976b) so that the method depends greatly on the analysis of constraints that arise at different stages of the processing. In this section and the next, we give a discursive account of the structures of our theory, and of the processes that employ them. The appendix describes a particular computer implementation of the theory, and gives an example of its application.

The 3-D model representation of shape

Our representation of 3-D shape is based on the idea of a stick figure, where each stick is the axis of a generalized cone (as defined above). For the purpose of this paper we shall limit ourselves still further, to regular cylinders in place of generalized cones. The basic element in the description of shape is called a *3-D model* and consists of:

(i) A *model axis*, which provides a very coarse specification of the general size and orientation of the shape.

(ii) A small number (possibly zero) of *component axes*. The component axes consist of a distinguished axis called the *principal axis* of the 3-D model, and a number of *auxiliary axes*. The dispositions of the auxiliary axes are defined relative to the principal axis, and that of the principal axis is defined relative to the model axis.

(iii) Associated with each axis is a shape description, which in the present restricted theory consists of the specification of a cylinder.

For example, the 3-D model for the overall shape of a human has six component axes in addition to the single model axis for the whole shape. The principal axis corresponds to the torso, and the five remaining component axes correspond to the head and limbs that are connected to it (see figure 6).

Although a single 3-D model is a simple structure, several may be combined to create a description of arbitrary depth and complexity. This is achieved by the *concatenation rule for 3-D models*, according to which a component axis of one 3-D model serves as the model axis axis of another. By combining 3-D models in this way, one can build up descriptions of a particular physical structure to whatever level of detail is required. Such a description is called the *3-D model representation* of a physical structure.

Figure 6 illustrates how model concatenation is used to create the 3-D model representation of a human shape, and it exhibits the hierarchy that concatenation induces. At the top level is the 3-D model for the overall human shape. As we saw above, this contains a single cylinder description of the overall shape (based on the model axis), and axes for each of the shape's six major components. The next level of detail contains 3-D models for each of these components. For example, the arm 3-D model consists of a model axis, which coincides with the arm auxiliary axis in the human 3-D model, and two component axes that correspond to the upper and lower arms. The hierarchy extends in similar fashion through 3-D models for the lower arm, hand, and finger, and each step is illustrated in figure 6. In this way, a 3-D model representation may be built to capture the geometry of a shape to whatever level of detail is required.

The underlying idea here is that in order to use the 3-D model

representation, the largest unit that has to be manipulated at any one time is small -- a single 3-D model -- yet the representation of any whole shape may be elaborate.

Thus the decomposition shown in figure 6 should be thought of not as the process of successively refining a single description, but instead as a representation system in which the balance between resolution and extent of description is flexible, and can change rapidly according to the needs of the moment. For instance, one cannot examine the fine detail of a hand without first reducing the scope of the examination to just the hand 3-D model. If the owner of the hand suddenly moves away, the focus of attention can quickly be shifted to a model near the top of the hierarchy in figure 6, since that is the level of description at which movements of the body as a whole are best described.

We have found the trade-off between scope and detail to be a useful one for the processes studied by our theory, because the information preserved at each level of the representation is just that needed by the processes that use this representation to interpret an image. For example in the analysis of a projected human figure, the orientation of the torso relative to the viewer is computed using information about the orientations and lengths of the limbs relative to the torso as they are projected in the image. This is just the information that is represented by the human 3-D model. The same holds true lower down, for 3-D models of smaller parts.

The important overall characteristics of the 3-D model representations for shape are: (1) the description provided by each 3-D model is quite simple while still possessing the shape information important to the processes that will use the 3-D model; (2) this technique produces descriptions that are canonical over variations that are not important in terms of recognition at least for the animal shapes examined here; and (3) the fidelity of the shape representations produced is easily improved, without changing existing 3-D models, by simply adding more 3-D models to the description to represent finer details.

The Structure of a 3-D Model

The important question for specifying the form of a single 3-D model is the manner in which the relative dispositions of its axes are specified. There are three candidate coordinate systems, viewer-centered, object-centered and local.

The viewer-centered system is the one in which comparisons with the image have eventually to be made. The image, and hence the projected axes computed from it are forced by the laws of optics to be based on a spherical coordinate system centered on the viewer. The difficulty with this system is that the descriptions produced depend upon the orientation of the viewed object relative to the viewer. For example a horse facing left produces an entirely different description from a horse facing right in the image. Minsky's multiple views representation accepts this difficulty and attempts to deal with each distinct view as a separate problem. A system based on the 3-D model idea requires that the underlying representation be independent of the viewing angle. This allows us to reject a viewer-centered coordinate system.

An object-centered coordinate system is one in which each axis that occurs anywhere in the 3-D model representation of an object be specified in a circumscribing frame of reference based, for example, on the top-level major axis of that

object. Such a system is a poor one for articulated shapes where axes are not rigidly connected. For example, if one moves an arm, one's fingers usually move with it. If each finger axis were represented solely by reference to the overall body axis, almost any movement of a high-level 3-D model in the 3-D model representation would render obsolete all information below that level in the hierarchy.

The natural choice is therefore to distribute the coordinate system making it local to each 3-D model. The position of the finger axis is specified relative to the hand, which in turn is specified relative to the arm, and this, to the torso. In order to discover the position of the finger relative to the torso, these intermediate relations need to be examined and interpreted. The crucial advantage of local 3-D coordinate systems is that they preserve the modularity of the 3-D model representation, which in turn enhances its flexibility. Using this scheme, it is easy to represent an elephant with one leg replaced by an automobile tyre, given 3-D models for an elephant and a tyre.

In order to specify the coordinate system for the 3-D model representation, it therefore suffices to describe how the spatial dispositions of the axes in a single 3-D model are determined relative to its principal axis. Figure 7 illustrates how this is accomplished. The length and orientation of an auxiliary axis is specified in spherical coordinates *(inclination, girdle, size)* or (θ, ϕ, r) where the principal axis itself defines the unit vector (0, 0, 1.0). The precise position of the auxiliary is determined by specifying its origin as a triple in cylindrical coordinates *(embedding-girdle, embedding-distance, position)* for (ϕ, r, z) about the principal axis. Once again the axis itself is (0, 0, 1.0). For both of these specifications, the direction of the zero girdle-angle, ϕ, has to be supplied in order to fix the angular rotation about the principal axis. The set

-(inclination, girdle, size, embedding-girdle, embedding-distance, position)

specifies the position of one cylinder relative to another, and it is called an *adjunct relation*.

Figure 7 shows the adjunct relation between the torso and left front leg of a cow. The leg starts at $(-100^{\circ}, 0.15, 0.8)$, that is, at the front end of the torso, displaced away from the axis of the torso by the torso's radius and located slightly ventral to the left side. From that point, the leg axis extends in a ventral direction about 2/3 of the torso's length $(90^{\circ}, 180^{\circ}, 0.66)$. Finally, the thickness of the leg is much less that that of the torso.

The angles and lengths that occur in these relations are represented in a system that specifies both a value and a tolerance (table I in the appendix). For example, it is possible to state that a particular axis (like the leg of a quadruped) is connected rather precisely at one end of the torso, is approximately vertical with about a ten degree tolerance out to the side (in girdle-angle), and a tolerance in inclination of about 70 degrees, which includes positions through which the leg normally swings.

The Image-Space Processor

We have seen how structural information about a shape is held by its 3-D model representation in a coordinate system that is essentially distributed. We also noticed that information from the image is expressed in a viewer-centered coordinate frame. These two systems have to be related, and the mechanism for accomplishing this is called the *image-space processor*.

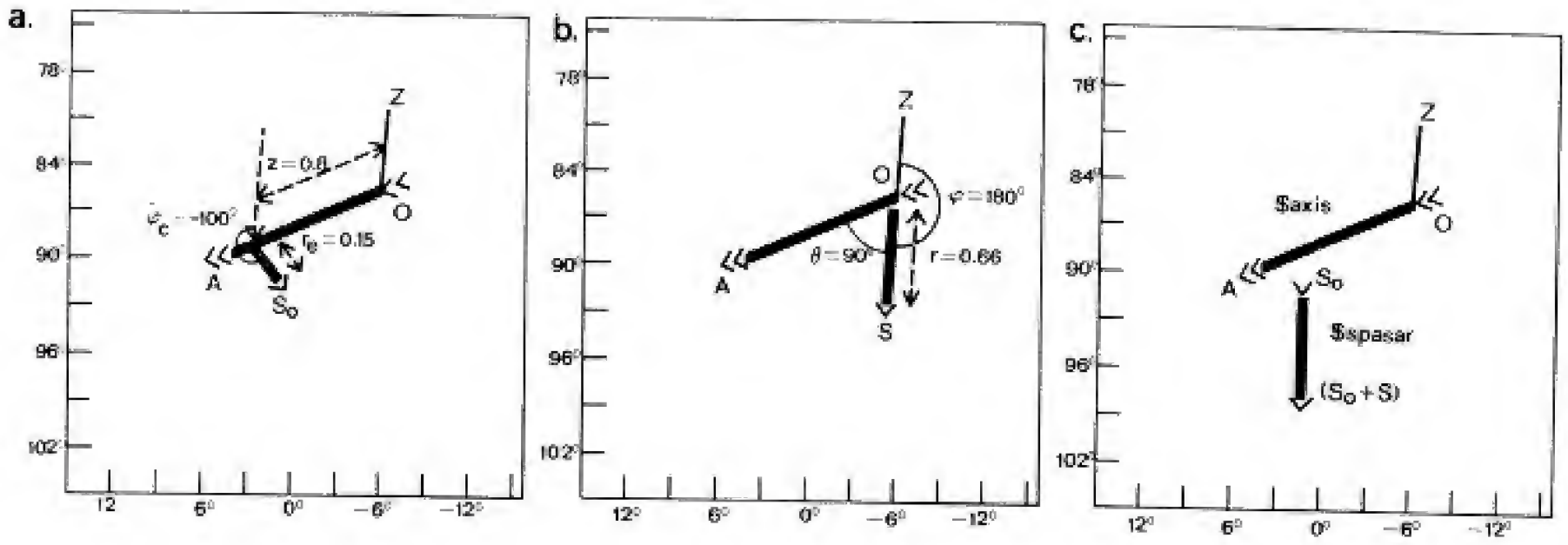


Figure 7. The spatial disposition of a new axis relative to the axis, $\overline{OA}$, in (a) above is specified by two vectors. One locates the origin, S_o relative to $\overline{OA}$, and the other specifies the new-axis' direction and length. The origin is specified in cylindrical coordinates *(embedding-girdle, embedding-distance, position)* for (ϕ_e, r_e, z), where $\overline{AO}$ defines the unit vector *(0, 0, 1)*. Figure 7a for example shows how to specify the origin of the left fore-leg of a cow relative to its torso axis. Next the direction and length of the left foreleg ($\overline{OS}$ in figure 7b) is specified in spherical coordinates *(inclination, girdle, size)* for (θ, ϕ, r), where $\overline{OA}$ again defines the unit vector *(0, 0, 1)*. In both notations, the girdle parameter is the angular rotation about $\overline{OA}$, relative to some "dorsal vector" that sets the zero girdle-angle. The sum of these two vectors locates the distal end of the left foreleg ($S+S_o$ in figure 7c). In the figures, the vector, $\overline{OZ}$, defines the zero girdle-angle, the ends of $\overline{OA}$ are marked with double-headed arrows, and in 7c the end points of the new axis are marked with single-headed arrows. This notational style will be used in subsequent figures to distinguish the *$axis* ($\overline{OA}$) and the *$spasar* ($\overline{S_oS+S_o}$).

Since our system for representing shape is based on 3-D models, each of which is simply a set of axes organized around a principle axis, the computational machinery needed in the image-space processor is very simple. It can be thought of as a tabular or simple arithmetic device that is able to maintain the representation of a distinguished vector, called the $axis, in viewer-centered spherical coordinates. In addition, the image-space processor can represent one movable vector called the $spasar (for space-arrow). The important point about the processor is that coordinates for the $spasar are available simultaneously in a frame centered on the viewer and in one centered on the $axis, so that specifying the $spasar in either frame makes it available in the other.

The $axis essentially defines a local coordinate system. It is specified by its two endpoints, and by one other point that defines the zero girdle-angle. The $spasar is defined by its two endpoints. Thus the image-space processor takes five points specifying the $spasar and $axis in the viewer centered system and produces an adjunct relation specifying the disposition of the $spasar relative to the $axis. The reverse transform, also computed by the image space processor, takes a specification of the $axis and a relation specifying the $spasar relative to the $axis, and produces the coordinates of the $spasar's end points in the viewer centered system. Since the viewer-centered system is expressed in spherical coordinates (θ, ϕ, r), predicted projections on the image may be obtained by simply ignoring the radial component r.

An example will help to clarify these points. If the orientation and location of the $axis relative to the viewer represents the torso axis of an imaginary horse and the appearance of its neck axis is required, the appropriate adjunct relation, giving the disposition of the neck axis relative to the torso axis, is read from the horse 3-D model and the image space processor is used to set the $spasar relative to the $axis as indicated by this relation. This computation produces the coordinates of the $spasar and thus the horse's neck axis in the viewer's reference frame and its projection is obtained by omitting the radial components.

In the simplest implementation of the image-space processor, the $axis is a passive element. Rotating it or translating it in the viewer's space-frame requires the use of the $spasar to compute its new coordinates. During recognition, two circumstances occur that cause one to move the $axis. Firstly, the orientation of a 3-D model is adjusted incrementally relative to the viewer until a disposition is found where the predictions from the 3-D model agree best with those obtained from the image. And secondly, when a piece of a 3-D model is to be examined in finer detail, one of the appendages of the model at the current level of study will become the principal axis for a more specialized model that deals with the fine structure of a sub-part. When shifting downwards to study the sub-part, the $axis and its implied reference frame has to be moved to the new principal axis. For example, when using the 3-D model for the overall structure of a man, the $axis will be bound to the torso. In order to move to a model for one of the arms, the $spasar must first be moved to that arm, and the $axis may then be transferred to the position computed by the $spasar.

The Catalogue of 3-D Models

The 3-D model representation of shape has been defined, and we have

seen in principle how the image-space processor relates the specifications found in a 3-D model representation to those being delivered from an image. The third major structure in the theory is a *catalogue* of stored 3-D models (see figure 8), from which individual 3-D models are freely selected and refined during the construction of the 3-D model representation for a given physical shape. The catalogue is indexed in various ways, so that incomplete shape information obtained during the analysis of an image causes a particular 3-D model to be selected; and this model, in turn, aids the further interpretation of the image by providing constraints on the possible dispositions of the axes found there.

The 3-D model catalogue may be thought of as a vocabulary of shape descriptions, and part of the process of recognition in our theory corresponds to the selection of increasingly specific 3-D models at each level of the 3-D model representation that is being built for the current image. Notice that making a 3-D model representation more specific by substituting increasingly specialized 3-D models within it is distinct from augmenting it with extra detail by adding new 3-D models to its fringes. In the first case, one might for example switch from an overall 3-D model for a quadruped to one for a horse; and in the second, one might add to the existing representation a 3-D model for a wart in the middle of one flank.

The 3-D model catalogue is organized in a hierarchy of increasing specificity. The topmost level contains the most undifferentiated description available, which is the 3-D model for a single cylinder. It is the top-level description of every shape in the catalogue. For this paper, we restrict the catalogue to those of a few animals, so at the next level of detail, there is a general quadruped shape, a primate shape, a bird-like shape, and various limbs. These schema are very general; for example, the quadruped shape specifies only that there are six appendages, with certain constraints on their positions and dispositions, but with only a very general specification of the types of limbs involved.

The 3-D model catalogue does not respect the difference between 3-D models for an object and its parts; its hierarchy simply traces lines of increasingly specialized description. Thus, 3-D models for the component parts of an object (legs, arms, ears, fingers, navels) are also arranged in the hierarchy of increasing specificity, while sharing the same top-level description of a single cylinder. For example, the hierarchy for a limb starts with the cylinder, next decomposes into two segments (like figure 8c), and each segment has its own subdivisions. In addition to this, the "general" (i.e. undifferentiated) limb 3-D model differentiates into forelimb and hindlimb, these into horse-forelimb, cow-forelimb, *etc.* At each level of specificity, a 3-D model has internal references to component sub-parts -- for example all limbs have upper and lower components -- and of course the upper-limb component of a horse-foreleg model differs from the upper component of a human-arm model.

The extent of this repertoire of shapes affects the efficiency of the computations for describing shapes presented to the system, but it does not limit one to them. For example, if presented with a favourable view of a horse like that in figure 4, a very limited system would be able to construct the description of its shape without the aid of a quadruped model using only single cylinder models, but it would take more time than if the quadruped model were available and used. Once the analysis of the shape in an image is

Figure 8. The 3-D model catalogue contains a repertoire of shapes organised from the general to the specific. It is consulted several times during the analysis of an image, and with its help a 3-D model representation of the viewed shape is constructed. At the top level is the most general model of all, a single cylinder. At the next level are models for general categories of shape; those listed here are for a quadruped, a primate, a bird and a limb. At the next level of differentiation, specific types of these general categories are represented. The constraints imposed, by using a model at one level in the catalogue to interpret an image, often give sufficient new information to enable one to select correctly a more specialized model. The organization exhibited in this figure is orthogonal to the organization depicted in figure 6.

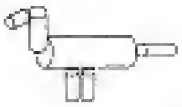
CYLINDER

LIMB

QUADRUPED

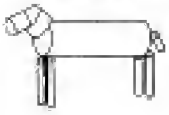
PRIMATE

BIRD

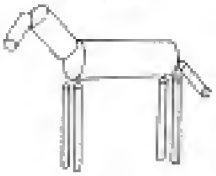


THICK-LIMB

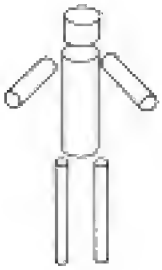
COW



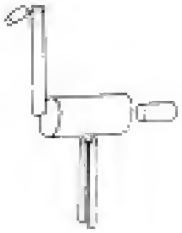
HORSE



HUMAN

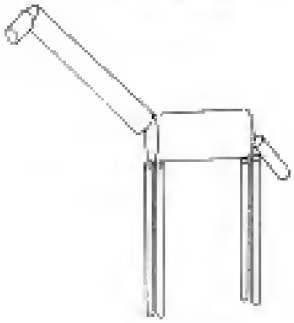


OSTRICH

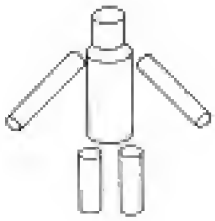


THIN-LIMB

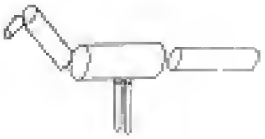
GIRAFFE



PRIMATE



DOVE



accomplished, the newly constructed 3-D models can be assigned to the catalogue as new models to be used to help interpret subsequent images. This step involves a considerable amount of indexing.

An important feature of the 3-D model catalogue is the extreme flexibility with which individual 3-D models may be used during the construction of a 3-D model representation for a given image. This is of course essential during the process of recognition, where the descriptions of the different parts of an object evolve independently to a certain extent. For example, one might at a particular instant be using a quadruped model, with rather general associated leg, neck and head models supporting the analysis. The constraints supplied by the head model allow a sufficient amount of new information to be obtained from the image so that the newly specialized description can be used to access the particular 3-D model for a horse-head directly *via* the catalogue's indexing mechanisms. This then allows the developing representation to be further specified both through improved specialization of the 3-D model selected for the whole animal's shape, and through improved specialization of the models for other components of the shape such as the head and legs.

III: The processes of the theory

We have seen how 3-D shapes are represented, and the mechanisms by which this representation is translated into quantities that may be measured from an image. We now turn to the more dynamical aspects of the theory, and these fall into two parts. First, how does one select an appropriate 3-D model given only the 2-D stick figure derived from an image? And second, having obtained a candidate 3-D model, how does its frame of reference come to be specified accurately relative to the viewer's? The basic strategy of our approach uses the principle of least commitment (Marr 1976b), which states that nothing should be done that may later have to be undone. At each stage, action is based on information and constraints that are reasonably certain, and is designed to produce new information and fresh constraints that will help to guide the analysis towards the desired goal.

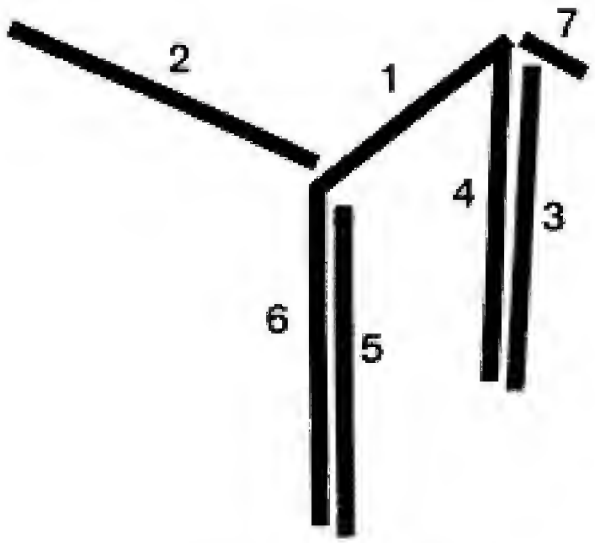
This part of the theory is only outlined; in fact it lies almost outside the 3-D representation module, since information from many other modules and interactions with them play an unavoidable role in the analysis of any but the simplest images.

The two homology problems

1: Accessing a suitable 3-D model

The first problem is how to obtain a suitable 3-D model. The database contains a large store of them, and we have to use information from the image to select one. The stored 3-D models range in specificity from the very general to the very particular (from a single cylinder to a giraffe), so that accessing the 3-D model database with a given set of features would in general cause the indexer to return many possible models. The principal of least commitment implies that one should never use a model that is more specific than current knowledge warrants, so it is inappropriate to index very specific models under very general attributes. Hence the access paths in the database behave more like a

a.



b.

3-D MODEL $QUADRUPED

COMPONENT AXES

1	-	$TORSO
2	-	$BUST
3	-	$LIMB
4	-	$LIMB
5	-	$LIMB
6	-	$LIMB
7	-	$TAIL

Figure 9. The homology problems. Previous visual processes deliver a datastructure like that exhibited in (a), where each axis is associated with a cylinder width, and the connectivity is explicitly available. The first homology problem is to select a suitable 3-D model from the catalogue. The result of the computations carried out here is the assignment of a *quadruped* 3-D model to this problem. Next, a homology must be established (so far as is possible) between the axes in the image and the component axes of the quadruped 3-D model. The result of this step is shown in (b). At this point the viewing angle is still unspecified, and only rather general information has been used to establish the homology with this unspecialized 3-D model.

decision tree than they would if every item were indexed independently. Once a general model like a quadruped has been retrieved and used to describe the image, it forms a local context through which more specialized features of that model can access more specialized 3-D models indexed under it.

Suppose that one is presented with a stick-figure image like that in figure 9. To begin with, nothing is known about the perspective from which the object is being viewed, so the initial 3-D model must be selected using information that is preserved by perspective transformations. Connectivity is not destroyed by perspective transformations, nor are quantities like the fractional distance down one axis at which another connects to it, unless the object is being viewed from very close by. Spurious connectivities can be introduced if one axis crosses in front of another and if the reason is not recognized lower down, but existing connections cannot be destroyed, only obscured. Hence in order to use connectivity information, when measuring which database items best match a given configuration set, unexplained errors of omission are treated much more seriously than unexplained errors of commission.

The second sort of information is girdle-angles, inclinations, and the relative lengths of axes. It is easier to take advantage of these later on, when the image-space processor has delivered at least partial results about the three-dimensional orientation relative to the viewer; but it is possible to do something with them early on. This comes about through weak, gross clues. For example if the 2-D length of the "neck" significantly exceeds the apparent length of a the "torso" in the image, and if the torso does not seem abnormally forshortened when compared with the length of the "legs", the image is likely to be a giraffe. In other words, lower bounds on the lengths of limbs can often be inferred, and are sometimes useful. Another important type of clue concerns major differences in the girdle-angles of two axes that are connected to a common one. For example, the neck and the tail often point in very different directions -- one up and one down -- and this obvious difference can usually be seen without a sophisticated 3-D analysis. In a pipe-cleaner animal, this very rough difference can help to determine which end of the animal is which.

The important point about the initial index access, and all subsequent accesses until an adequate description has been built, is that the newly selected model is used to structure information that is already available and is instrumental in obtaining further shape information from the image. This added information is then used to select a more specific model, and the process repeats itself until enough information is gathered for the purpose at hand.

The path to a 3-D model is not always direct. When an important stick in the stick-figure is foreshortened and component shapes are insufficient for determining the 3-D model, other kinds of strategies are needed. An interesting example is a water-pail (see figure 2). When seen from the side, the image of a pail segments naturally into its generalized cylinder description in which the pail is represented as the slice of a cone and the axis is vertical (figure 2c). If one looks down from above however, one essentially sees two circles joined by the sloping sides. The principal axis of the pail would appear as a point from this perspective (figure 2d), and if the pail's handle were missing or only vaguely defined in the image, there would be no strong component clues to work with.

In order to access the correct 3-D model despite these obfuscations, some idea of depth has to be introduced into the analysis *before* addressing the 3-D model index can be successful. In the case of the pail, some process has to realise that the two circles might be separated in depth, and that if they are, they could be separated by a considerable distance. The clues that signal this in monocular images include radial symmetry and nuances of shadow and highlight, which leads us to expect that much of the analysis of lighting and shadow can influence the processing at exactly this stage of recognition. We think of the computations that take place here as deploying the Espasar to construct from the image a primary 3-D model, that consists at first of an axis in depth whose circumscribing surface is bounded by the two visible circles, and to which extra details - like hollowness, the closure of one end of this surface by an orthogonal plane, and possibly the addition of a cross-strut to account for the handle - are added. At some point during the construction of this description, the indexer is successful at finding a match with some near antecedent of the bucket 3-D model in the catalogue. If an "unconventional view" becomes a common view, it would become profitable to index the appropriate 3-D model under the special features that obtain for that view.

2: Matching the image to a model

Once a 3-D model has been selected, its component axes must be paired with sticks in the stick-figure image. Since the ways in which a 3-D model is selected vary considerably, the association between these elements is not always automatic. Often, some of the associations will remain ambiguous. For example, imagine the silhouette of a horse from the side; the legs are easily identified but the left and right forelegs cannot be distinguished without further information. What is important in many cases is that a particular stick from the image is one of the legs, since the legs are roughly parallel and it is their orientation rather than their specific identity that is important for computing the figure's shape.

The information available for making these associations increases as the processing proceeds. Initially, positional information along the principal axis of the stick figure is depended upon most heavily. Often, clues that are available at this stage include the relative thicknesses of the shapes round the stick axes (the neck of a horse is much thicker than the legs), and the decompositions of component sticks (the tail and legs of a horse may be roughly straight, but the bust has two components that always make a large angle with one another). Symmetry or repetition can also be important for disambiguating the components of a stick figure. For example the legs of a horse are all the same thickness, are roughly parallel, and because of this have roughly the same length in the stick-figure image, distinguishing them from the tail. Also the legs and tail are usually on one side of the torso while the bust extends to the other side in the image of a horse. Collectively, such clues are often sufficient to disambiguate the major components of a 3-D model.

Relaxation

The final part of the theory assumes that the image has been described by a 3-D model with which a homology has been established, and describes how the model

Figure 10. Relaxing a stored model onto the stick-figure derived from the image. Once a 3-D model has been selected and associations have been made between the axes of the model and the sticks computed from the image, the approximate orientation of the model relative to the viewer is computed *via* a hill climbing algorithm using the image space processor. This process is carried out with the $axis positioned so that its projection coincides with the stick associated with the model's principal axis (as indicated by the double-headed arrows above). With this arrangement, the appropriateness of a proposed principal axis orientation can be judged by using the $spasar to compare the consequent projections of the model's limbs with the associated sticks in the image. The $axis can be rotated in two dimensions without moving its projection away from its assigned stick in the image. It can be dipped toward or away from the viewer and it can be rotated about its own axis. In the figures above, dark lines indicate sticks computed from the image and light lines are projections computed using the $spasar. The top sequence (a), (b), (c) shows the projected axis of the quadruped model for different rotations about the $axis while its ends are equidistant from the viewer. In the lower sequence (d), (e), (f) the tail end of the $axis is moved slightly farther away from the viewer.

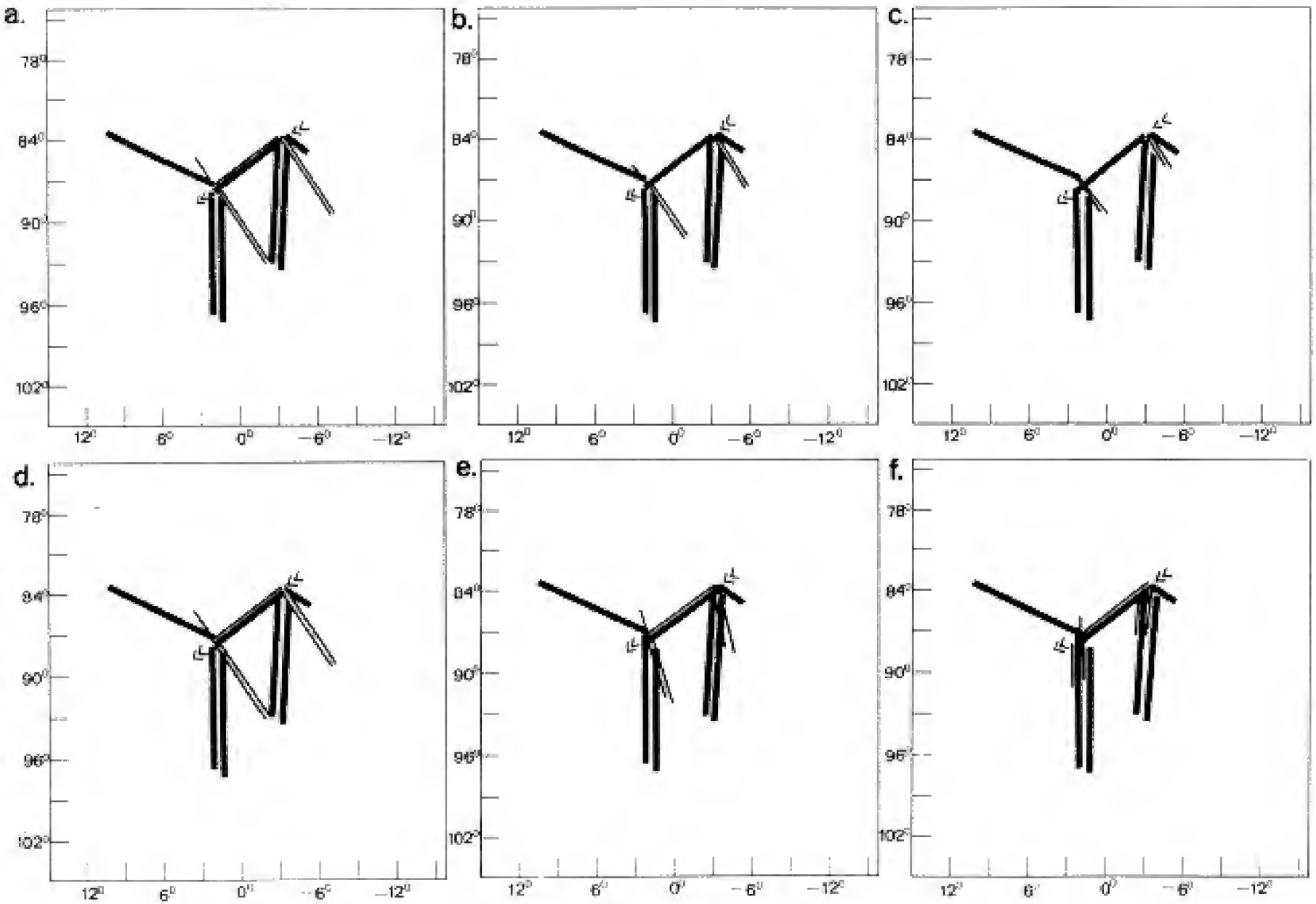
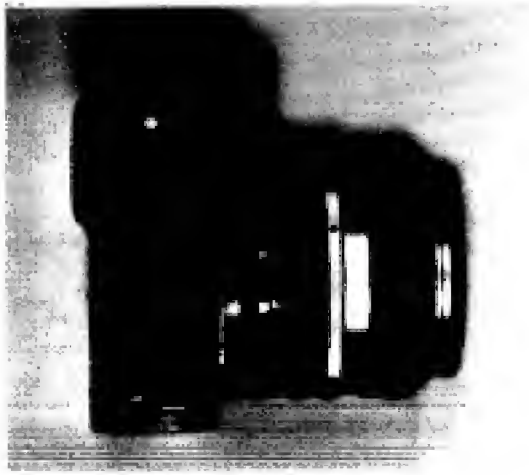

a.
b.
c.
d.
e.
f.
78°
84°
90°
96°
102°
12°
6°
0°
−6°
−12°

Figure 11. Views of an object in which an important axis is foreshortened are surprisingly common. From only one of these views of a camera (b), may its two main axes be recovered straightforwardly from the image. Figures (d) through (f) show how this happens, by displaying the axes for each of the views (a) - (c) within a line drawing of the overall shape. Views (a) and (c) fall into the same class as the top view of a water-pail (figure 2b). According to the theory, the class of such views provides a rigorous definition of the intuitive notion of an "unconventional" view (Warrington & Taylor 1973).

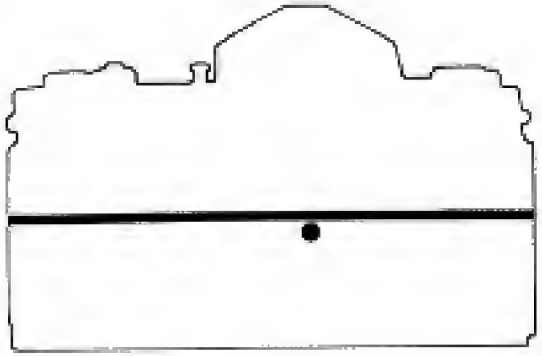
a.

b.

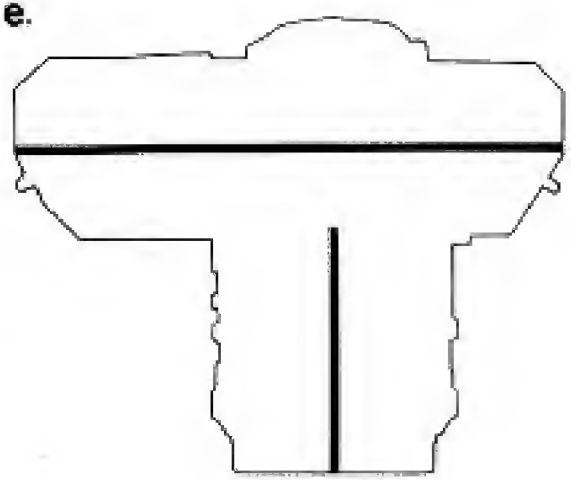
c.



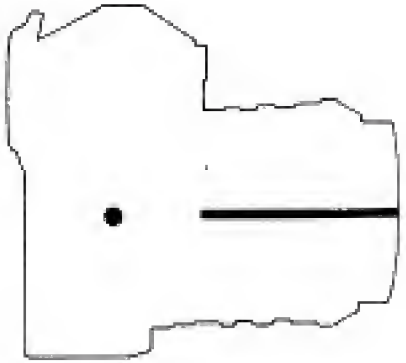
d.



e.



f.



comes to possess the appropriate 3-D orientation relative to the viewer. This is accomplished by an incremental hill-climbing procedure which uses the image-space processor and information in the 3-D model to match the model to the axes derived from the image.

The basic idea here is to use the image-space processor to compute the discrepancy between a given 3-D model orientation and the constraints imposed by the stick-figure image. The $axis is set to an arbitrary initial orientation (slightly approaching or receding from the viewer, based perhaps on shading cues) so that its projection is parallel to the axis of the stick-figure image. Two degrees of freedom are left unconstrained at this point, the dip of the $axis out of the image plane towards or away from the viewer, and the unit vector associated with the $axis which determines the rotation about the $axis of the object's local coordinate system (see figure 10). From a given disposition, the discrepancy between the 3-D model's projected component axes and the corresponding sticks of the image can be computed using the image-space processor, and their sum gives an indication of the goodness of fit of this particular orientation of the $axis. A simple incremental hill-climbing technique may now be used, that varies the dip of the $axis and the rotation about it until a suitably good fit is found. Further discussion of the process illustrated in figure 10 may be found in the appendix.

This technique is incomplete as it stands, since the orthogonal projection of a stick figure looks the same regardless of whether its head is nearer the viewer than its tail. For animals like a horse, this ambiguity may be resolved by noticing whether the forelegs or the hindlegs are shorter. For less familiar objects, obscuration or context clues (what the object is on or in) are probably necessary to disambiguate the two possibilities.

Finally, comparison with the angles of the image are only a partial source of error information in the hill-climbing computation. Used alone, they would make the computed disposition of the $axis too sensitive to slight variations in the dispositions of the component axes in the image. We therefore include in the error calculation discrepancies between the dip of the $spasar away from the viewer and the dip computed from the image using perspective information (does the circumscribing cylinder thicken at the nearer end as it should?), and length information (for this orientation of the $spasar is its projection too long or too short compared with the image?). Our grasp of this part of the theory is adequate only for simple images, and we shall develop it further elsewhere.

IV: Discussion

The discussion falls naturally into two parts, one concerned specifically with vision, and the other with the organization of information in a wider sense.

1: 3-D representation theory

There are five main points to our theory. They are:

(1) The 3-D disposition of an object is represented primarily by a stick-figure configuration, where each stick stands for one or more axes in the object's generalized cone representation. (2) This configuration is described by a loosely hierarchical assertional database, called a 3-D model representation. Use of this database is extremely free and flexible, and it can support levels of description that cover the spectrum from very coarse to very fine detail. It also satisfies the principle of graceful degradation, which states that partial information

should yield partial results.
(3) In order to be useful, this database has to be interpreted through an (essentially) analogue mechanism, called the image-space processor. In its minimal implementation, this processor can be thought of as maintaining the representation of one vector in a local space-frame.
(4) The image-space processor's instruction set is small. Its most important features are:
(a) the ability to interpret an adjunct relation between the $axis and the $spasar; and
(b) the ability to relate object-centered coordinates to a viewer-centered frame of reference.
(5) The image-space processor can deliver information about the lengths and orientations of the appearance of the $axis and $spasar. These help the system to "rotate" its model into the correct 3-D disposition relative to the viewer.

The immediate and most accessible prediction that follows from the theory concerns the characterization of Warrington & Taylor's (1973) "unconventional" views. According to our theory, the most difficult views to handle are those in which an important axis is foreshortened, since in these cases straightforward segmentation fails to recover them from the image. We therefore predict that these are the views that Warrington & Taylor would label unconventional, and on which their patients will fail most easily. Such views are by no means uncommon, and figures 2 and II contain two familiar examples.

It is hard but not impossible to derive detailed neurophysiological predictions from the theory, particularly predictions about the likely implementation of the image-space processor (Nishihara, in preparation). There are however several general points about the theory that lead us to take it seriously as a model for psychology, and which therefore encourage us to derive more detailed predictions. They are:
(1) Pipe-cleaner animals are almost as easily recognizable as are line-drawings of animals, despite their very abstract relation to the original. This would not be surprising if pipe-cleaner animals were in some sense extracted from the image during the normal course of its interpretation (as our theory asserts), but it would be surprising if not.
(2) The loosely hierarchical structure of our 3-D models has many computational advantages that are almost bound to be shared by the psychological representation, even if the psychological representation is otherwise very different. The advantages include a variable level of detail in the 3-D model system, and the flexibility with which different 3-D models may be accessed and combined to form new models. If a system has 3-D models for a horse and for a man, it will be able to build the description of a centaur.
(3) An important part of the theory is the simplicity of the image-space processor. The only requirements are that it be able to manipulate one vector in a space-frame, and relate the specification in that frame to one in the viewer-centered frame. By using the stick-figure representation, the essentials of the spatial organization of a shape may be manipulated at very low computational cost.
(4) The mechanisms of the theory can handle 3-D shapes, and so are inherently powerful enough to describe 2-D patterns, such as the configuration of features on a face. The only requirement is that such patterns should be described relative to axes that are constructed within them, since the structure of a 3-D model depends on specifying positions in this way.

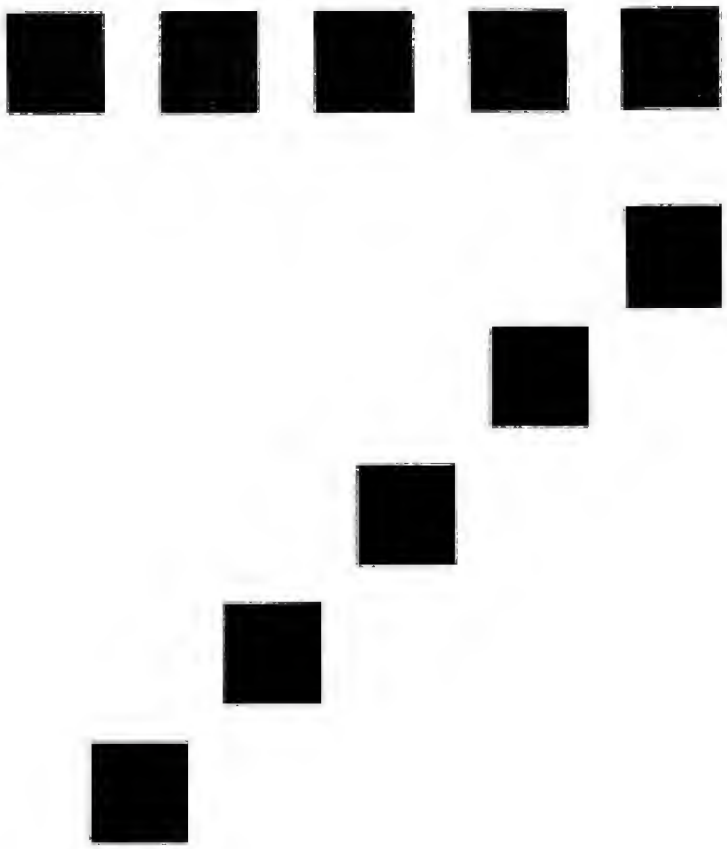


Figure 12. This figure, taken from Minsky & Papert (1972), illustrates the influence of an axis on the description of a figure. In one row, the shapes are seen as squares, and in the other, as diamonds. The establishing of axes in a 2-dimensional figure is important for our theory, since it determines how the description of a 2-D configuration is constructed. This figure is the 2-D analog of figure 5, since it establishes a reasonable case that one precondition for using our theory as a psychological model - namely the computation of axes during the analysis of 2-D patterns - is satisfied by our visual systems.

It is therefore important for the theory that axes be established early in our perception of 2-D figures. Figure 12 provides positive evidence on this point. In the top row, the shapes are seen as squares, whereas along the diagonal, they are seen as diamonds. The diagonal axis is therefore being constructed during the analysis of this pattern; it influences, and therefore probably precedes, the description of the shapes of the local elements.

(5) The theory has been implemented and works well for simple images (see the appendix).

Mental rotation experiments

In 1971, Shepard & Metzler (1971) created a set of images by rotating and reflecting simple objects made of cubes (figure 13). They found that the time taken to decide whether two such images were of identical objects, rather than objects that differed by a reflexion, varied linearly with the angle through which one object must be rotated in 3-space to become aligned with the other. This finding revived interest in "mental imagery" and in analogue processes in perception (Cooper & Shepard (1973), Metzler & Shepard (1974), Shepard (1975)). In addition, Kosslyn (1975) has published evidence for an analogue component to the processes that interpret mainly two-dimensional structures, like faces and maps.

The significance of such experiments is controversial (but not the results). Part of the reason for the controversy seems to have been some difficulty in seeing how an "analogue" process could benefit the computations that underlie perception and recognition. We believe that the present theory shows a way in which such a mechanism could be useful. It asserts that there is indeed an analogue component to the process, namely the image-space processor, and that it operates on the sticks in a 3-D model. The linearity that Shepard *et al.* regard as significant is however not a deep consequence of our theory, merely the signature of one particularly simple way of implementing it. In the language of Marr & Poggio (1976b), the linearity is a consequence more of the mechanisms that are used than of the underlying nature of the computation.

Broadly speaking, if our theory is taken as a psychological model, it predicts three stages in the assignment of 3-D orientation to views that are not unconventional. The stages are: (a) A startup period, during which the axes are obtained from the image, the 3-D model database is accessed, and the two homology problems are solved. (b) An incremental process, during which the stored 3-D model is relaxed onto the axes being delivered from the image. This process uses the principal axis together with the two or three other most suitable ones, and in its simplest incremental implementation the time for relaxation will vary roughly linearly with the 3-D angle through which the stored model's space-frame is rotated. (c) Finally, when the best 3-D orientation has been found, the remaining axes in the model are bound to the image, and fine adjustments made to their positions and sizes.

The same computational theory certainly has other equally viable implementations that do not exhibit a linear dependence on the angle. In one of these implementations, the angle through which the model's frame is rotated at each increment is half the angle between its present position and the currently predicted final state. In this implementation, the time to settle would vary with approximately the logarithm of the 3-D angle. Such a system does not have so starkly simple an image-space processor as the linear

one, but its requirements are still modest relative to what a digital electronic computer can provide. It must also be borne in mind that unless the subject is very familiar with the objects being recognized, the interaction between the image, the image-space processor, and the 3-D model database may be extended and complex. In such cases, any linear dependence on angle could be masked completely by the process of accessing successively more detailed 3-D models. This is particularly true if the subject is presented with an unconventional view of an unusual or unfamiliar object, an expectation that suggests several experiments.

If one bears this caveat in mind, however, only one of the findings reviewed by Shepard (1975, item 14 page 100) is unexpected. It comes from Cooper & Shepard (1973b condition O), who showed that advance information giving the orientation but not the identity of the object to be presented is not sufficient to enable subjects to prepare for it. One might have expected that subjects could rotate their $axis to the appropriate orientation, and leave it there to be bound to the principal axis of a 3-D model when the image was presented. In order to incorporate this finding, we would need to assume (for example) that the image space processor cannot be run unless bound to a 3-D model (even if only of an arrow), and that whenever the $axis is rebound to a radically new 3-D model, the image-space processor is reset. There are some other grounds for wanting this. The space-frame in the image-space processor needs more than one direction to define it, and trying to construct a space-frame round a given vector can lead to problems if the 3-D model is not simple. Secondly, in the real world, one rarely sees two objects at the same point in the field of view. Therefore, to change to a new 3-D model almost always requires a change in the direction of gaze. In order to compensate for this in a minimal implementation, the $axis and $spasar would have to be set to axes in the starting frame, in order to carry out the primary rotations that allow for the angle of gaze. These arguments are however weaker than the arguments that support the rest of the theory.

Before we leave the discussion of the visual aspects of the theory, it is appropriate to note that the 3-D model representation is not without its disadvantages. Firstly, it is based on the structural axes of a shape, and some attempt at extracting them must be made before the mechanisms of the theory can be invoked. To do so requires a great deal of pre-processing of the image, and the theory associated with this is only beginning to be worked out (see Marr & Poggio 1976b for a brief review). For views in which a structural axis is foreshortened, this pre-processing may be completely unable to deliver the correct axes. On such views, a system that operates according to the present overall theory will be severely disadvantaged. It is not clear whether other methods exist that would be more successful.

Finally, the criticism about the absence of uniqueness, that we made of Baumgart's system for the representation of shapes by polyhedral approximation, sometimes applies to the generalized cone representation. For example, consider a doorway. The natural axis of most doors is vertical, because they are higher than they are wide. This is not always true, however, and it is perfectly possibly to represent a doorway by an axis parallel to the width of the door, or even one parallel to its thickness. For most purposes, there is little difference between using the height and using the width as the principal axis, but using the thickness may introduce an important new way of looking at the space the

door occupies, since when arranged in this direction, the $spasar carries information about the direction that is involved in passing through it. In other words, the analysis and use of holes may depend to a considerable extent on using the $spasar to define what "through" a hole means. Moreover, we feel that many of the problems of representing and manipulating the space immediately around the viewer can be handled conveniently and efficiently using a mechanism like the image-space processor.

2: Broader issues concerning the representation of knowledge

Following the tradition of Bartlett (1932), Minsky (1975) observed that the "chunks" of reasoning, language, memory and perception ought to be larger and more structured than most theories in artificial intelligence and psychology allow. This idea is much more attractive than it is easy to realise, and two factors can be identified as mainly responsible for the difficulty. The first is what are the chunks? To answer it, one must know how to represent a piece of knowledge for the purpose at hand, and much work in artificial intelligence is devoted to asking this question in different domains. Sometimes it is answered with conspicuous success (Moses (1971 MACSYMA), Shortliffe (1976 MYCIN), Duffield *et al.* (1969, DENDRAL), Sussman & Stallman (1975 EL)).

The second factor is the question of flexibility. If all one's knowledge resides in canned chunks, little room remains for variations in a scenario that are inevitable in each of its real-world instances. This factor causes particular difficulties in domains that are ambitiously near to real-world situations, like Schank's (1975) restaurant scenario. Its effect is to leave these scenarios unable to deal with irregularities.

In the present theory, we propose that the central description of shape is based on the 3-D model representation. The desired flexibility is achieved by modularity within the representation, which allows 3-D models to be combined as the image dictates, and by using the 3-D model catalogue more as an aid to building the current description than as a set of inviolate subunits that must be assembled unchanged in a rigid way.

The other point that we believe may be important about the theory is the way it embodies Minsky's assertion, that the overall structure of a situation or shape is of importance to the way its details are recognized and their organization represented. The key idea here is the use of coarse overall descriptions of a shape to help extract new information from the image, which in turn enables the 3-D models involved in its description to be specialized further so that yet more can be read from the image. Thus, 3-D models for the overall structure of a shape set up a context of spatial constraints, between otherwise unrelated axes in the image, which then allow specific local "deductions" to refine the details -- possibly causing the overall description to be abandoned. This process is directly analogous to the situation in Sussman & Stallman's (1975) program for understanding electronic circuits, where a "high-level" description like "voltage-divider" becomes attached to part of a circuit, relating components by local laws that are special and informative, and which allow constraints on the behaviour of that part to be stated accurately and concisely. In these two domains, these phenomena seem to capture the essence of what makes Minsky's (1975) article so stimulating, although we feel that the interplay between different levels of description, which forms a crucial part of the computation, has yet to receive a satisfactory

general formulation. In any case, the important feature of these two examples is that they specify precisely the information contained in the high-level descriptions. Discussions that consider only possible implementation *mechanisms* (frames, semantic networks, property lists, Conniver methods, actors etc.) are not useful for deciding how information should be represented in a fresh domain.

The explicit nature of these high-level organizing structures (the quadruped, the voltage-divider) stands in sharp contrast to methods based on cooperative phenomena, like the stereopsis theory of Marr & Poggio (1976a), in which the higher-level "holistic" organizing structure of the computation remains an implicit, not an explicit, aspect of the network by which it is implemented.

There may be an interesting connection between the specific database organization that is required by our theory, and a recent study of human semantic memory. The organization that makes it possible to carry out the construction of a gradually more specific 3-D model representation is the ordering of the 3-D model catalogue by increasingly specific shape. Thus the access sequence for 3-D models during the recognition of (say) a mallard-shape would often be approximately:

small-blob-shape --> bird-shape --> duck-shape --> mallard-shape.

This provides an interesting functional basis for structuring the 3-D model catalogue according to rules very similar to those exhibited by Warrington (1975), in a recent and ingenious study of the structure of semantic memory.

Finally, we feel that a simple mechanism along the lines of the image-space processor would be of great benefit to a motor control system. At some level, a motor system must have access to a representation of body-space in which distances, directions and trajectories are computed and stored in a form closely related to what visual information can provide. Yet to execute a motor action, the commands must eventually be couched in terms of lengths, tensions and joint angles. A mechanism along the lines of the image-space processor could provide a link between the two, at low computational cost.

Figure 14. This figure exhibits the information contained in 3-D models currently listed in the restricted 3-D model catalogue used by our present implementation. Each 3-D model, referenced by its $name, has an associated width and a list of relations among its component axes. This list of relations specifies the relative spatial dispositions of the components, and indicates a 3-D model for each one. The accompanying stick figures show the appearance of these components relative to one another from a particular vantage point.

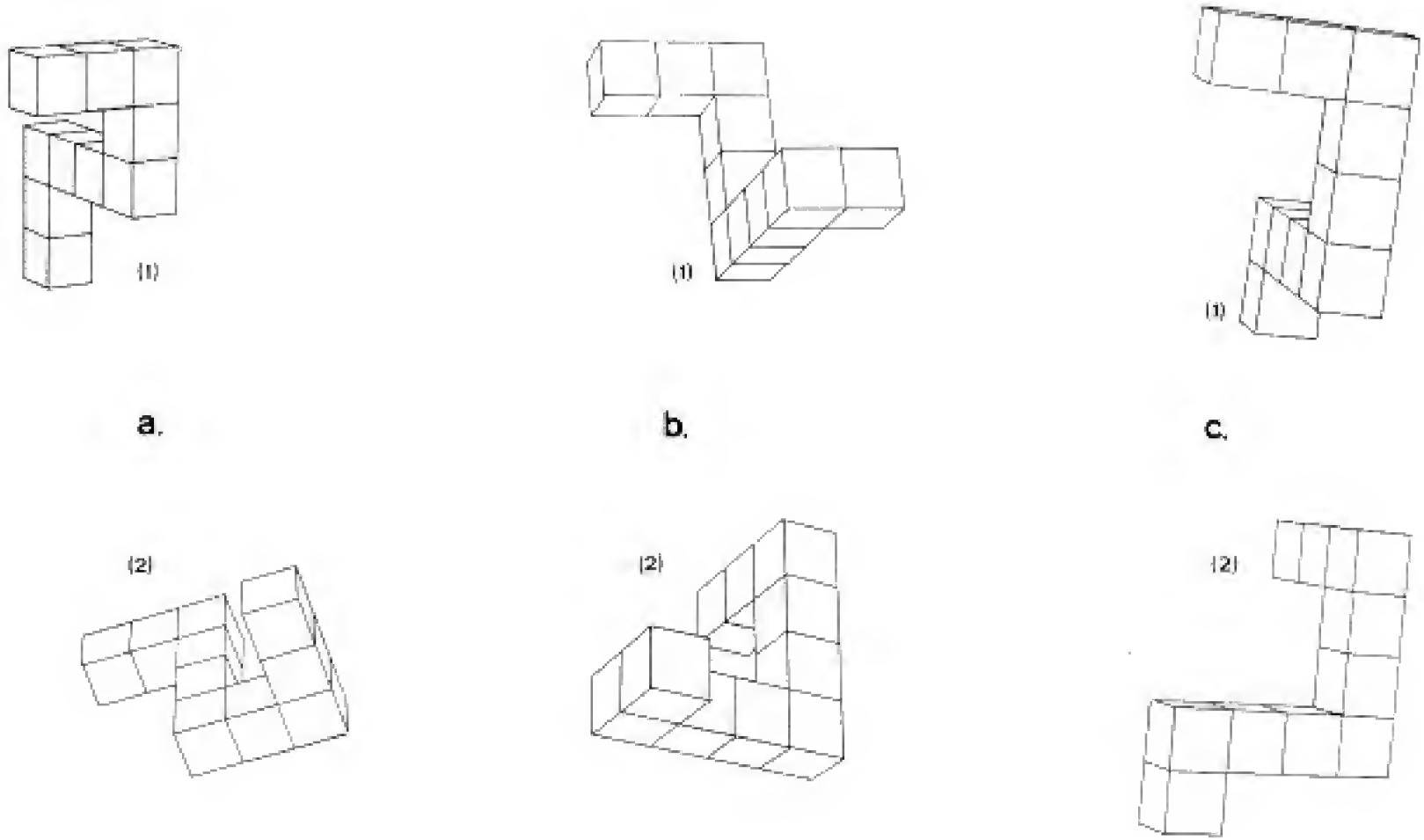


Figure 13. Examples of pairs of perspective line drawings presented to the subjects of Shepard & Metzler's (1971) experiments on mental rotation. (A) A "same" pair, which differs by an 80 degree rotation in the picture plane; (B) a "same" pair which differs by an 80 degree rotation in depth; (C) a "different" pair, which cannot be brought into congruence by any rotation. The time taken to decide whether a pair is the "same" varies linearly with the (3-D) angle by which one must be rotated to be brought into correspondence with the other, (reconstructed after figure 1 of Shepard & Metzler, 1971).

```
$CYLINDER
  WIDTH:  N

$QUADRUPED
  WIDTH:  N
  RELATIONS:
        ($QUADRUPED $TORSO POS N GIRD N INCL N EMBG N EMBD S SIZE N)
        ($TORSO $BUST POS S GIRD N INCL W EMBG N EMBD S SIZE E)
        ($TORSO $LIMB POS N GIRD S INCL W EMBG E EMBD E SIZE N)
        ($TORSO $LIMB POS N GIRD S INCL W EMBG W EMBD E SIZE N)
        ($TORSO $LIMB POS S GIRD S INCL W EMBG E EMBD E SIZE N)
        ($TORSO $LIMB POS S GIRD S INCL W EMBG W EMBD E SIZE N)
        ($TORSO $TAIL POS N GIRD S INCL W EMBG N EMBD S SIZE E)

$PRIMATE
  WIDTH:  N
  RELATIONS:
        ($PRIMATE $TORSO POS W GIRD N INCL N EMBG N EMBD S SIZE N)
        ($TORSO $HEAD POS S GIRD N INCL N EMBG N EMBD S SIZE E)
        ($TORSO $LIMB POS S GIRD W INCL W EMBG W EMBD N SIZE N)
        ($TORSO $LIMB POS S GIRD E INCL W EMBG E EMBD N SIZE N)
        ($TORSO $LIMB POS N GIRD N INCL S EMBG W EMBD N SIZE N)
        ($TORSO $LIMB POS N GIRD N INCL S EMBG E EMBD N SIZE N)

$BIRD
  WIDTH:  N
  RELATIONS:
        ($BIRD $TORSO POS W GIRD N INCL N EMBG N EMBD S SIZE N)
        ($TORSO $BUST POS S GIRD N INCL W EMBG N EMBD S SIZE E)
        ($TORSO $LIMB POS W GIRD S INCL W EMBG W EMBD E SIZE E)
        ($TORSO $LIMB POS W GIRD S INCL W EMBG E EMBD E SIZE E)
        ($TORSO $TAIL POS N GIRD S INCL S EMBG N EMBD S SIZE E)

$TORSO
  WIDTH:  N

$BUST
  WIDTH:  N
  RELATIONS:
        ($BUST $NECK POS N GIRD N INCL N EMBG N EMBD S SIZE N)
        ($NECK $HEAD POS S GIRD S INCL W EMBG S EMBD E SIZE E)

$TAIL
  WIDTH:  E

$LIMB
  WIDTH:  E
  RELATIONS:
        ($LIMB $UPPER-LIMB POS N GIRD N INCL N EMBG N EMBD S SIZE E)
        ($UPPER-LIMB $LOWER-LIMB POS S GIRD N INCL N EMBG N EMBD E SIZE N)

$UPPER-LIMB
  WIDTH:  E

$LOWER-LIMB
  WIDTH:  E
  RELATIONS:
        ($LOWER-LIMB $FORELIMB POS N GIRD N INCL N EMBG N EMBD S SIZE N)
        ($FORELIMB $PAW POS S GIRD N INCL N EMBG N EMBD E SIZE E)

$PAW
  WIDTH:  NN
  RELATIONS:
        ($PAW $PALM POS NN GIRD NN INCL NN EMBG NN EMBD NN SIZE EN)
        ($PALM $FINGER POS WS GIRD WW INCL WW EMBG WW EMBD NW SIZE EN)
        ($PALM $FINGER POS SS GIRD NN INCL NN EMBG WW EMBD NW SIZE NN)
        ($PALM $FINGER POS SS GIRD NN INCL NN EMBG WW EMBD EN SIZE NN)
        ($PALM $FINGER POS SS GIRD NN INCL NN EMBG EE EMBD EN SIZE NN)
        ($PALM $FINGER POS SS GIRD NN INCL NN EMBG EE EMBD NW SIZE NN)
```

```
$HORSE
  WIDTH:  NW
  RELATIONS:
        (&HORSE $TORSO POS NN GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $BUST POS SS GIRD NN INCL NW EMBG NN EMBD EE SIZE EN)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG EE EMBD EE SIZE NN)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG WW EMBD EE SIZE NN)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG EE EMBD EE SIZE NN)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG WW EMBD EE SIZE NN)
        (&TORSO $TAIL POS NN GIRD SS INCL WS EMBG NN EMBD SS SIZE EN)
```

```
$COW
  WIDTH:  NW
  RELATIONS:
        (&COW $TORSO POS NN GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $BUST POS SS GIRD NN INCL NW EMBG NN EMBD EE SIZE SE)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG EE EMBD EE SIZE EN)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG WW EMBD EE SIZE EN)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG EE EMBD EE SIZE EN)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG WW EMBD EE SIZE EN)
        (&TORSO $TAIL POS NN GIRD SS INCL WS EMBG NN EMBD SS SIZE SE)
```

```
$GIRAFFE
  WIDTH:  NW
  RELATIONS:
        (&GIRAFFE $TORSO POS NN GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $BUST POS SS GIRD NN INCL NW EMBG NN EMBD SE SIZE NW)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG EE EMBD EE SIZE NW)
        (&TORSO $LIMB POS NN GIRD SS INCL WW EMBG WW EMBD EE SIZE NW)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG EE EMBD EE SIZE NW)
        (&TORSO $LIMB POS SS GIRD SS INCL WW EMBG WW EMBD EE SIZE NW)
        (&TORSO $TAIL POS NN GIRD SS INCL WS EMBG NN EMBD SS SIZE EN)
```

```
$HUMAN
  WIDTH:  EN
  RELATIONS:
        (&HUMAN $TORSO POS WW GIRD NN INCL NN EMBG NN EMBD SS SIZE EN)
        (&TORSO $HEAD POS SS GIRD NN INCL NN EMBG NN EMBD SS SIZE EE)
        (&TORSO $LIMB POS SS GIRD WW INCL WS EMBG WW EMBD NN SIZE EN)
        (&TORSO $LIMB POS SS GIRD EE INCL WS EMBG EE EMBD NN SIZE EN)
        (&TORSO $LIMB POS NN GIRD NN INCL SS EMBG WW EMBD NN SIZE NN)
        (&TORSO $LIMB POS NN GIRD NN INCL SS EMBG EE EMBD NN SIZE NN)
```

```
$MONKEY
  WIDTH:  EN
  RELATIONS:
        (&MONKEY $TORSO POS WW GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $HEAD POS SS GIRD NN INCL NN EMBG NN EMBD SS SIZE EE)
        (&TORSO $LIMB POS SS GIRD WW INCL WS EMBG WW EMBD NN SIZE NN)
        (&TORSO $LIMB POS SS GIRD EE INCL WS EMBG EE EMBD NN SIZE NN)
        (&TORSO $LIMB POS NN GIRD NN INCL SS EMBG WW EMBD NN SIZE EN)
        (&TORSO $LIMB POS NN GIRD NN INCL SS EMBG EE EMBD NN SIZE EN)
```

```
$OSTRICH
  WIDTH:  NN
  RELATIONS:
        (&BIRD $TORSO POS WW GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $BUST POS SS GIRD NN INCL WW EMBG NN EMBD SS SIZE NN)
        (&TORSO $LIMB POS WW GIRD SS INCL WW EMBG WW EMBD EE SIZE NN)
        (&TORSO $LIMB POS WW GIRD SS INCL WW EMBG EE EMBD EE SIZE NN)
        (&TORSO $TAIL POS NN GIRD SS INCL SS EMBG NN EMBD SS SIZE EE)
```

```
$DOVE
  WIDTH:  NN
  RELATIONS:
        (&BIRD $TORSO POS WW GIRD NN INCL NN EMBG NN EMBD SS SIZE NN)
        (&TORSO $BUST POS SS GIRD NN INCL NW EMBG NN EMBD SS SIZE EN)
        (&TORSO $LIMB POS WW GIRD SS INCL WW EMBG WW EMBD EE SIZE EN)
        (&TORSO $LIMB POS WW GIRD SS INCL WW EMBG EE EMBD EE SIZE EN)
        (&TORSO $TAIL POS NN GIRD SS INCL SS EMBG NN EMBD SS SIZE NN)
```

Appendix: an implementation of the theory

In an ideal situation, a theory of visual information processing would consist entirely of well-defined, circumscribed results, accompanied by proofs of existence and uniqueness, with enough background to show that the results obtained are in fact those that are important for visual information processing. Marr (1976a) labelled such theories Type I. If this were always the case, there would be considerable interest in, but no serious need for implementing the theory on a computer. In the present state of the art one is rarely so fortunate, since even when an individual module can be given a Type I theory, the interactions between it and other modules cannot be satisfactorily analyzed until the other modules have themselves been graced with Type I theories. This is to some extent the situation here. The core of the present theory is of Type I -- the 3-D representation is well-defined, and the image-space processor is precisely formulated -- but in analyzing the interactions between the 3-D module itself and other visual or non-visual processes that pass it clues, many different kinds of information have to be taken into account.

It is therefore important to implement a theory such as this, and writing its implementation has proved an important technique for clarifying our ideas and testing different approaches to carrying out a process. For example, algorithms for accessing the 3-D model catalogue are peripheral to the 3-D representation theory, but they are essential to a program that implements it. In our present implementation these algorithms are quite primitive, because the main focus of our attention has been on the image-space processor and on relaxation mechanisms. We have postponed the development of a more psychological candidate for the catalogue indexing mechanisms until the important access paths into the catalogue are more clearly defined.

The purpose of this appendix on the implementation is therefore to clarify some of the concepts peripheral to the theory, and to lend substance to the notions set out in it by exhibiting them at work. We make no claims that the implementation we describe here is optimal, and it is certainly not unique.

Database conventions

Each 3-D model in our current implementation is organized around a special name such as *$quadruped*, *$limb*, or even *$0001*, and we call these names *$names (dollar-names)*. Each $name specifies a memory location in the computer where the various shape informations associated with the particular 3-D model are stored, and the $name is used to reference that information. Many of the $names in this appendix, such as *$quadruped* or *$limb*, are mnemonics for the shape the associated 3-D model represents. This clarifies the presentation, but is of no other significance.

Figure I4 exhibits the information contained in 3-D models currently listed in the restricted 3-D model catalogue used by our present implementation. Each 3-D model, referenced by its $name, has an associated width and a list of relations among its component axes. This list of relations specifies the relative spatial dispositions of the components, and indicates a 3-D model for each one. For example the first relation in the *$primate* template is

($primate $torso pos W gird N incl N embg N embd S size N),

TABLE 1a

Representation of angle and of position

Directions and angles that occur in an adjunct relation are expressed in a vocabulary of symbols that define a value and a tolerance. The longer the symbol, the more accurately it specifies a value. Tables 1a & b define the values and tolerances of all symbols that occur in the figures.

	N	W	S	E	
	0.25	0.75	*	*	upper limit
POS	0.0	0.5	1.0	*	center
	*	0.25	0.75	*	lower limit
	45.0	135.0	-135.0	-45.0	upper limit
GIRD	0.0	90.0	180.0	-90.0	center
	-45.0	45.0	135.0	-135.0	lower limit
	45.0	135.0	*	*	upper limit
INCL	0.0	90.0	180.0	*	center
	*	45.0	135.0	*	lower limit
	45.0	135.0	-135.0	-45.0	upper limit
EMBG	0.0	90.0	180.0	-90.0	center
	-45.0	45.0	135.0	-135.0	lower limit

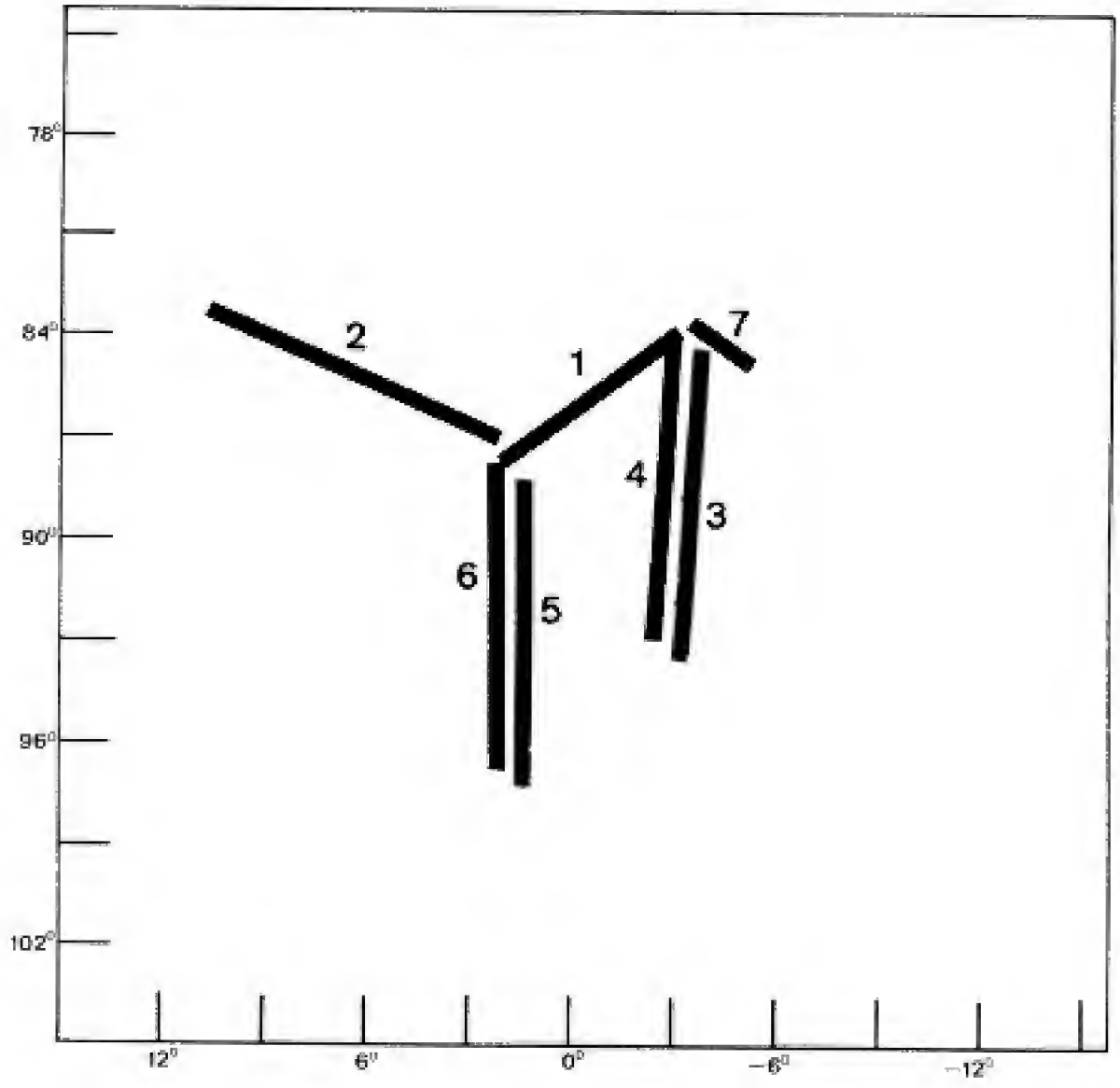
	S	E	N	W	
	0.04	0.12	0.32	0.89	upper limit
EMBD	0.02	0.07	0.2	0.54	center
	0.0	0.04	0.12	0.32	lower limit
	0.22	0.6	1.64	4.48	upper limit
SIZE	0.13	0.36	1.0	2.71	center
	0.0	0.22	0.6	1.64	lower limit
	0.08	0.24	0.65	1.79	upper limit
WIDTH	0.05	0.14	0.4	1.08	center
	0.0	0.08	0.24	0.65	lower limit

TABLE 1b

	NN	NW	WW	WS	SS	SE	EE	EN	
	0.12	0.37	0.62	0.87	*	*	*	*	upper limit
POS	0.0	0.25	0.5	0.75	1.0	*	*	*	center
	*	0.12	0.37	0.62	0.87	*	*	*	lower limit
	22.5	67.5	112.5	157.5	-157.5	-112.5	-67.5	-22.5	upper limit
GIRD	0.0	45.0	90.0	135.0	180.0	-135.0	-90.0	-45.0	center
	-22.5	22.5	67.5	112.5	157.5	-157.5	-112.5	-67.5	lower limit
	22.5	67.5	112.5	157.5	*	*	*	*	upper limit
INCL	0.0	45.0	90.0	135.0	180.0	*	*	*	center
	*	22.5	67.5	112.5	157.5	*	*	*	lower limit
	22.5	67.5	112.5	157.5	-157.5	-112.5	-67.5	-22.5	upper limit
EMBC	0.0	45.0	90.0	135.0	180.0	-135.0	-90.0	-45.0	center
	-22.5	22.5	67.5	112.5	157.5	-157.5	-112.5	-67.5	lower limit

	SS	SE	EE	EN	NN	NW	WW	WS	
	0.03	0.05	0.09	0.15	0.25	0.42	0.69	1.15	upper limit
EMBD	0.02	0.04	0.07	0.12	0.2	0.32	0.54	0.89	center
	0.0	0.03	0.05	0.09	0.15	0.25	0.42	0.69	lower limit
	0.17	0.28	0.47	0.77	1.28	2.11	3.48	5.75	upper limit
SIZE	0.13	0.22	0.36	0.6	1.0	1.64	2.71	4.48	center
	0.0	0.17	0.28	0.47	0.77	1.28	2.11	3.49	lower limit
	0.06	0.11	0.18	0.31	0.51	0.84	1.39	2.3	upper limit
WIDTH	0.05	0.08	0.14	0.24	0.4	0.65	1.08	1.79	center
	0.0	0.06	0.11	0.18	0.31	0.51	0.84	1.39	lower limit

Figure 15. The information supplied by earlier visual processes consists of a collection of two-dimensional stick descriptions, together with information about the thickness associated with each, and their connectivity. The example shown here has been simplified to include only the sticks for the top level 3-D model. The dollar name *$0000* is the reference for a new 3-D model that will eventually represent the shape of this stick figure. The FIGURE property of *$0000* relates the organization found in the image to the structure required of a 3-D model, indicating syntactically that stick 0 is the top-level single axis decription of the overall shape, stick 1 is the principal component of this shape, and sticks 2 through 7 are its auxiliary axes. The table specifies the angular locations of the end-points of each of these sticks in a viewer-centred coordinate system, along with their thicknesses.



$0000
FIGURE: (0 (1) (2) (3) (4) (5) (6) (7))
PACKET: TRUE

stick#	end a			end b		
	θ	φ	width	θ	φ	width
0	83.9	-3.1	4.9	87.5	2.0	5.1
1	83.7	-3.1	3.0	87.3	2.0	3.1
2	86.9	2.0	5.2	83.3	10.3	5.4
3	83.9	-3.6	1.8	92.8	-2.9	1.7
4	83.5	-2.7	1.8	92.4	-2.1	1.7
5	87.6	1.6	1.9	96.6	1.6	1.8
6	87.1	2.5	1.9	96.1	2.4	1.8
7	83.5	-3.2	0.6	84.7	-4.9	0.6

note: all values are in degrees

which specifies the disposition of the *$torso* cylinder relative to the *$primate* cylinder. The *$primate* cylinder is the single cylinder representation of the whole primate shape, and the *$torso* is one of its component cylinders. Notice that *$torso* is the dollar-name of another 3-D model. This is how the concatenation rule between 3-D models is implemented.

The other information in this relation consists of attribute-value pairs, such as *pos W* and *gird N*. These two pairs specify the position of the *$torso* cylinder along the *$primate* axis to be *W* (which means in the middle, between 0.25 and 0.75); and the girdle-angle to be *N* (which means within 45 degrees of 0). The symbols *N*, *W*, *S*, *NN*, *NNNW etc.* specify directions and tolerances, the longer symbols specifying a direction more precisely than the shorter ones. Table I defines the values and tolerances of all the symbols used here.

The *$torso* is the principal axis of the *$primate* 3-D model, and the remaining relations held in the model specify the dispositions of the auxiliary axes relative to it. Here, there are six auxiliary axes, the *$head*, *$tail*, and four *$limbs*.

The form of the input

The information supplied by earlier visual processes consists of a collection of two-dimensional stick descriptions, together with information about the thickness associated with each, and their connectivity. Figure 4 in the main text was obtained from a grey-level image using the techniques described by Marr (1976b), and it illustrates how information about axes may be obtained from an image. Figure 15 shows an example that has been simplified by omitting the embedding relations. The dollar name *$0000* is the reference for a new 3-D model that will eventually represent the shape of this stick figure. The FIGURE property of *$0000* relates the organization found in the image to the structure required of a 3-D model. The information held here indicates that stick 0 is the top-level single axis model for the overall shape, stick 1 is the principal axis for the first elaborated 3-D model, and sticks 2 through 7 are the auxiliary axes for this model. In a more detailed example, these auxiliary axes would themselves decompose into substructures. It might be the case, for example, that stick 2 (the figure's bust) decomposed into two component sticks, corresponding to the neck and head. If this added detail had been included in the input data, (2) would be replaced by (2 (8) (9)) in the FIGURE property of $0000.

Homology and the primary catalogue access

The first step in the 3-D analysis of such a figure is to select an appropriate 3-D model from the catalogue, and to match it to the incoming stick figure (the two homology problems). This is done by computing estimates of the adjuncts between the principal axis of the stick figure and its auxiliaries, and then selecting that 3-D model whose adjunct relations are most similar to the estimated ones.

If the radial coordinates of the end points of the sticks in figure 15 were known, the image-space processor could be used to compute the required adjunct relations directly. They are not; but it turns out that useful relations can be obtained this way by first assuming that all the radial distances are the same, which is equivalent to interpreting

the image as if all its sticks lay perpendicular to the line of sight. This is the starting configuration for the 3-D model axes, and as the processing continues, better values for the radial coordinates will be established.

Figure 16 shows the result of translating this initial configuration into adjunct relations via the image-space processor. Note that low resolution symbols have been used in the computed relations, and that new 3-D models for each auxiliary axis have been created. The girdle-angles depend upon the particular choice of zero girdle direction, which is initially arbitrary. The only important thing about them now is that some of the relations have *gird E*, and some have *gird W*, which correspond "above" and "below" the principal axis. The position parameter at this point is reasonably accurate, up to possible reversal if the wrong end of the principal axis has been taken as the zero end.

Only positional values along the principal axis provide direct help for selecting a 3-D model from the catalogue, and even this is subject to a possible polarity error. The remaining parameters do however provide indirect help, even though they may be severely distorted by the *a priori* assumption that the sticks are coplanar. For example, although the inclinations, girdles and sizes may themselves be incorrect, certain interrelations among them will be preserved; the neck will have a girdle angle of opposite sign to the legs', and the legs will all have similar inclinations, girdles and sizes because they are roughly parallel. This information is sufficient to select the *$quadruped* model from the *$quadruped*, *$bird*, *$primate* and various *$limb* models, and the relations in *$0000* can now be associated with the relations held in *$quadruped*. This information is inserted into the newly formed 3-D models under their TEMPLATE properties as shown in figure 17.

Relaxation

The next step is to use information in the *$quadruped* model to compute better estimates for the radii, beginning with the principal axis, stick 1. Figure 10 of the main text shows how our program achieves this. A hill-climbing algorithm is used, where the parameters to be adjusted are the radial coordinate of one of stick 1's end points, and the zero girdle direction of the *$axis*, which lies along stick 1. The *$axis* represents the current attempt at matching the *$torso* to stick 1, and as the *$axis* is incrementally rotated, the goodness of fit of the 3-D model is computed by placing the *$spasar* successively on the *$torso* relations of the *$quadruped*, and accumulating a similarity score between the *$spasar's* end-points and the associated sticks in the image. In the top row of figure 10, the end points of the *$axis* are equidistant from the viewer for three successive orientations of the zero-girdle direction. The "appearance" of the *$quadruped* is computed one axis at a time, and is shown in lighter lines in the figure. The effect of rotating about the *$axis* does not significantly improve the fit. In the bottom row, the radial value of the end of the *$axis* has been improved, and now rotation about the *$axis* leads to a good alignment. This sets a new estimate for the radial coordinates of stick 1, and now, the image-space processor is used to set new estimates for the radial components of the remaining sticks, based on relations stored in the *$quadruped*.

```
$0000
  RELATIONS:
        ($0000 $0001 POS N GIRD E INCL N EMBG N EMBD S SIZE N)
        ($0001 $0002 POS S GIRD E INCL N EMBG S EMBD S SIZE N)
        ($0001 $0003 POS N GIRD W INCL N EMBG N EMBD S SIZE N)
        ($0001 $0004 POS N GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0005 POS S GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0006 POS S GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0007 POS N GIRD W INCL W EMBG S EMBD S SIZE E)
  WIDTH: W
  FIGURE: (0 (1) (2) (3) (4) (5) (6) (7))
  PACKET:  TRUE
  TEMPLATE:  $CYLINDER

$0001
  WIDTH: N
  FIGURE:    (1)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0002
  WIDTH: N
  FIGURE:    (2)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0003
  WIDTH: E
  FIGURE:    (3)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0004
  WIDTH: E
  FIGURE:    (4)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0005
  WIDTH: E
  FIGURE:    (5)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0006
  WIDTH: E
  FIGURE:    (6)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER

$0007
  WIDTH: E
  FIGURE:    (7)
  PACKET:  $0000
  TEMPLATE:  $CYLINDER
```

Figure 16. The first step in the processing of the input information is the computation of a model-centred description of the sticks. Radial information is required in order to use the image space processor to compute this decription but it is not supplied in the input. It turns out however that useful relations can be obtained by assuming that the radial distances to the end points of the sticks are the same. This is equivalent to assuming that all the sticks of the image are in a plane perpendicular to the line of sight. The result of translating this initial configuration into adjunct relations via the image-space processor using this assumption is shown here. Note that low resolution symbols have been used in the computed relations, and that new 3-D models for each auxiliary axis have been created. The girdle-angles depend upon the particular choice of zero girdle direction, which is arbitrary initially. The only important thing about them now is that some of the relations have *gird E*, and some have *gird W*, which correspond to "above" and "below" the principal axis. The position parameter at this point is reasonably accurate, up to possible reversal if the wrong end of the principal axis has been taken as the zero end.

```
$0000
  RELATIONS:
        ($0000 $0001 POS N GIRD E INCL N EMBG N EMBD S SIZE N)
        ($0001 $0002 POS S GIRD E INCL N EMBG S EMBD S SIZE N)
        ($0001 $0003 POS N GIRD W INCL N EMBG N EMBD S SIZE N)
        ($0001 $0004 POS N GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0005 POS S GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0006 POS S GIRD W INCL N EMBG S EMBD S SIZE N)
        ($0001 $0007 POS N GIRD W INCL W EMBG S EMBD S SIZE E)
  WIDTH: W
  FIGURE: (0 (1) (2) (3) (4) (5) (6) (7))
  PACKET:  TRUE
  TEMPLATE:  $QUADRUPED

$0001
  WIDTH: N
  FIGURE:    (1)
  PACKET:  $0000
  TEMPLATE:  $TORSO

$0002
  WIDTH: N
  FIGURE:    (2)
  PACKET:  $0000
  TEMPLATE:  $BUST

$0003
  WIDTH: E
  FIGURE:    (3)
  PACKET:  $0000
  TEMPLATE:  $LIMB

$0004
  WIDTH: E
  FIGURE:    (4)
  PACKET:  $0000
  TEMPLATE:  $LIMB

$0005
  WIDTH: E
  FIGURE:    (5)
  PACKET:  $0000
  TEMPLATE:  $LIMB

$0006
  WIDTH: E
  FIGURE:    (6)
  PACKET:  $0000
  TEMPLATE:  $LIMB

$0007
  WIDTH: E
  FIGURE:    (7)
  PACKET:  $0000
  TEMPLATE:  $TAIL
```

Figure 17. The positional distribution of the adjunct sticks (three appendages at each end of the principal axis), along with similarity relations derived from the *gird*, *incl*, *size*, and *width* parameters (four appendages, two on each end are very similar while a remaining one is very different), are used to select a general 3-D model from the 3-D model catalogue. In this case the *$quadruped* model was selected as indicated by the template property listed under *$0000*. The second homology is also carried out here assigning template properties to the components of *$0000* and relating adjunct relations in *$0000* to adjunct relations in *$quadruped* (these latter assignments are not depicted here).

```
$0000
  RELATIONS:
        ($0000 $0001 POS NN GIRD NN INCL NN EMBG SS EMBD SS SIZE NN)
        ($0001 $0002 POS SS GIRD NN INCL NW EMBG NN EMBD NN SIZE NW)
        ($0001 $0003 POS NN GIRD SS INCL WW EMBG EN EMBD SS SIZE NW)
        ($0001 $0004 POS NN GIRD SS INCL WW EMBG WS EMBD WS SIZE NW)
        ($0001 $0005 POS SS GIRD SS INCL WW EMBG EN EMBD NN SIZE NW)
        ($0001 $0006 POS SS GIRD SS INCL WW EMBG WS EMBD WS SIZE NW)
        ($0001 $0007 POS NN GIRD SS INCL WS EMBG EN EMBD SS SIZE EN)
  WIDTH: NW
  FIGURE: (0 (1) (2) (3) (4) (5) (6) (7))
  PACKET:  TRUE
  TEMPLATE:  $QUADRUPED
```

Figure 18. We see here the state of *$0000* just after the completion of the relaxation process depicted in figure 10. The adjunct relations have been recomputed by the image space processor using symbols with a slightly higher level of resolution.

```
$0000
  RELATIONS:
       ($0000 $0001 POS NN GIRD NN INCL NN EMBG SS EMBD SS SIZE NN)
       ($0001 $0002 POS SS GIRD NN INCL NW EMBG NN EMBD NN SIZE NW)
       ($0001 $0003 POS NN GIRD SS INCL WW EMBG EN EMBD SS SIZE NW)
       ($0001 $0004 POS NN GIRD SS INCL WW EMBG WS EMBD WS SIZE NW)
       ($0001 $0005 POS SS GIRD SS INCL WW EMBG EN EMBD NN SIZE NW)
       ($0001 $0006 POS SS GIRD SS INCL WW EMBG WS EMBD WS SIZE NW)
       ($0001 $0007 POS NN GIRD SS INCL WS EMBG EN EMBD SS SIZE EN)
  WIDTH: NW
  FIGURE: (0 (1) (2) (3) (4) (5) (6) (7))
  PACKET: TRUE
  TEMPLATE:  $GIRAFFE
```

Figure 19. The adjunct relations in figure 18 are used again to access a 3-D model from the 3-D model catalogue. This access results in the selection of the *$giraffe* 3-D model, based largely on the lengths of the neck and legs relative to the torso, and the first stage of recognition is complete.

Secondary catalogue access and recognition

Having found the 3-D model orientation that achieves the best fit, we can now compute a new set of adjunct relations for *$0000*, the model that is being built. These are shown in figure 18. Notice that we are now using symbols with a slightly higher level of resolution. The 3-D model catalogue can now be accessed in search of a more specific shape. This access results in the selection of the *$giraffe* 3-D model, based largely on the lengths of the neck and legs relative to the torso, and the first stage of recognition is complete. The final state of the 3-D model is shown in figure 19.

Acknowledgements: We thank Drew McDermott, Tomaso Poggio, and Kent Stevens for valuable criticism, and Karen Prendergast for preparing the drawings. This article describes work reported in M. I. T. A. I. Lab. Memo 341, and it was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-75-C-0634.

References

Adrian, E. D. 1941 Afferent discharges to the cerebral cortex from peripheral sense organs. *J. Physiol. (Lond.), 100*, 159-191.

Agin, G. J. 1972 Representation and description of curved objects. *Stanford A. I. Memo 173*.

Allman, J. M., Kaas, J. H., Lane, R. H. & Miezin, F. M. 1972 A representation of the visual field in the inferior nucleus of the pulvinar in the owl monkey (*Aotus trivirgatus*). *Brain Research, 40*, 291-302.

Allman, J. M., Kaas, J. H. & Lane, R. H. 1973 The middle temporal visual area (MT) in the bushbaby, *Galago senegalensis*. *Brain Research, 57*, 197-202.

Allman, J. M. & Kaas, J. H. 1974a The organization of the second visual area (V-II) in the owl monkey: a second order transformation of the visual hemifield. *Brain Research, 76*, 247-265.

Allman, J. M. & Kaas, J. H. 1974b A visual area adjoining the second visual area (V-II) on the medial wall of parieto-occipital cortex of the owl monkey (*Aotus trivirgatus*). *Anat. Rec., 178*, 297-8.

Allman, J. M. & Kaas, J. H. 1974c A crescent-shaped cortical visual area surrounding the middle temporal area (MT) in the owl monkey (*Aotus trivirgatus*). *Brain Research, 81*, 199-213.

Binford, T. O. 1971 Visual perception by computer. Presented to the IEEE Conference on Systems and Control, Miami, in December 1971.

Blum, H. 1973 Biological shape and visual science, (part I). *J. theor. Biol., 38*, 205-287.

Brodmann, K. 1909 *Vergleichende Lokalisationlehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: J. A. Barth.

Cajal, S. Ramon y, 1911 *Histologie du systeme nerveux de l'homme et des vertebres*. 2 vols. Paris: Norbert Maloine.

Cooper, L. A. & Shepard, R. N. 1973 a The time required to prepare for a rotated stimulus. *Memory and Cognition, 1*, 246-250.

Cooper, L. A. & Shepard, R. N. 1973 b Chronometric studies of the rotation of mental images. In: W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Critchley, M. 1953 *The parietal lobes*. London: Edward Arnold.

Duffield, A. M. et al. 1969 Applications of artificial intelligence for chemical inference, II: Interpretation of low-resolution mass spectra of ketones. *J. Am. Chem. Soc., 91*, 2977-2981.

Hollerbach, J. M. 1975 Hierarchical shape description by selection and modification of prototypes. *M. I. T. Master's Thesis, to appear as M. I. T. A. I. Lab. TR-346.*

Kosslyn, S. M. 1975 Information representation in visual images. *Cognitive Psychology, 7*, 341-370.

Luria, A. R. 1970 *Traumatic aphasia*. The Hague: Moulton.

Marr, D. 1976a Artificial Intelligence -- a personal view. *M. I. T. A. I. Lab. Memo 355*

Marr, D. 1976b Early processing of visual information. *Phil. Trans. Roy. Soc. B.*, (in the press).

Marr, D. 1976c. Analysis of occluding contour. *M. I. T. A. I. Lab. Memo 372.*

Marr, D. & Poggio, T. 1976a Cooperative computation of stereo disparity. *Science, (submitted for publication).* Also available as *M. I. T. A. I. Lab. Memo 364.*

Marr, D. & Poggio, T. 1976b From understanding computation to understanding neural circuitry. In *The Visual Field: Psychophysics and Neurophysiology. Neurosciences Research Program Bulletin*, E. Poeppel *et al.*, Eds. (in the press).

Metzler, J. & Shepard, R. N. 1974 Transformational studies of the internal representation of three-dimensional objects. In: *Theories of cognitive psychology: The Loyola Symposium*, Ed. R. Solso. Hillsdale, N. J.: Lawrence Erlbaum Assoc.

Minsky, M. 1975 A framework for representing knowledge. In: *The psychology of computer vision*, Ed. P. H. Winston, pp 211-277. New York: McGraw-Hill.

Minsky, M. & Papert, S. 1972 Artificial intelligence progress report. *M. I. T. A. I. Lab. Memo 252.*

Moses, J. 1974 MACSYMA -- the fifth year. *SIGSAM Bulletin, ACM, 8*, 105-110. See also *The MACSYMA reference manual*, M. I. T. Laboratory for Computer Science, 545 Technology Square, Cambridge, Mass. 02139.

Nevatia, R. 1974 Structured descriptions of complex curved objects for recognition and visual memory. *Stanford A. I. Memo 250.*

Schank, R. C. 1975 *Conceptual information processing.* New York: Elsevier.

Shepard, R. N. 1975 Form, formation, and transformation of internal representations. In: *Information processing and cognition: The Loyola Symposium,* Ed. R. Solso, pp 87-122. Hillsdale, N. J.: Lawrence Erlbaum Assoc.

Shepard, R. N. & Metzler, J. 1971 Mental rotation of three-dimensional objects. *Science, 171,* 701-703.

Shortliffe, E. H. 1976 *Computer-Based Medical Consultations: MYCIN,* New York: American Elsevier Publishing Company, Inc.

Street, R. F. A. 1931 A Gestalt completion test: A study of a cross-section of intellect. In: *Teachers College Contributions to Education, No. 481.* New York: Teachers College, Columbia University.

Ullman, S. 1976 Structure from motion. *(M. I. T. Ph. D. Thesis in preparation).*

Vatan, P. & Marr, D. 1976. Algorithms for the decomposition of a contour. In perparation.

Vinken, P. J. & Bruyn, G. W. 1969 Eds. *Handbook of Clinical Neurology: Vol. 2, Localization in Clinical Neurology.* (In association with A. Biemond). Amsterdam: North Holland Publishing Co.

Warrington, E. K. & Taylor, A. M. 1973 The contribution of the right parietal lobe to object recognition. *Cortex,* 9, 152-164.
Also remarks made by E. K. W. in a lecture given on Oct. 26th 1973 at the M. I. T. Psychology Department.

Warrington, E. K. 1975 The selective impairment of semantic memory. *Quart. J. exp. Psychol.,* 27, 635-657.

Zeki, S. M. 1971 Cortical projections from two prestriate areas in the monkey. *Brain Research, 34,* 19-35.

Zeki, S. M. 1973 Colour coding in rhesus monkey prestriate cortex. *Brain Research, 53,* 422-427.